A Comparative Analysis of Linear and Machine Learning Models for Used Car Price Prediction

Jihoon Choi[1], Dae-young Kim[1], and Wei Sun[1]

[1]Michigan State University, College of Natural Sciences

December 13, 2024

Abstract

Purpose - This study explores predictive models for estimating used car prices, addressing challenges in price prediction by analyzing both linear and machine learning approaches. It aims to understand the complex relationships between automotive characteristics and market values, filling a critical gap in automotive market analytics.

Design/Methodology/Approach - The study compares four predictive models: Ordinary Least Squares (OLS) regression, ElasticNet regularization, Random Forest, and XGBoost. A dataset sourced from Kaggle, containing 188,000 samples with information on car prices and features such as mileage, model year, and accident history, was analyzed. Model performance was evaluated using metrics including Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and the coefficient of determination (R^2).

Findings - Machine learning models, particularly Random Forest and XGBoost, outperformed linear models (OLS and ElasticNet) in capturing non-linear relationships, demonstrating superior accuracy and practical utility. Ensemble methods proved most effective in addressing complex interactions between vehicle features. Key predictors included mileage, model year, and engine specifications, reflecting their strong influence on pricing.

Originality/Value - This research provides a novel comparative framework for used car price prediction, highlighting the advantages of machine learning techniques such as Random Forest and XGBoost. The findings offer practical insights for model selection and feature importance, contributing to better decision-making in the automotive industry. Future research may incorporate external market trends to further enhance predictive accuracy.

Keywords: used car prices; predictive modeling; Random Forest; XGBoost; machine learning; pricing strategies

1 Introduction

Predicting used car prices based on vehicle features has become increasingly important in the automotive industry, serving as a crucial tool for various stakeholders in the market. Accurate price predictions not only facilitate efficient transactions between buyers and sellers but also provide valuable insights into consumer preferences and market trends. Understanding the relationship between specific vehicle features and their market values enables dealerships, manufacturers, and financial institutions to make data-driven decisions about inventory management, pricing strategies, and market positioning (Gutierrez, 2024).

The complexity of this prediction task stems from the numerous factors that influence a vehicle's value (Narayana *et al.*, 2021), including mileage, model year, fuel type, and accident history. These features often interact in complex ways, making it challenging to develop models that accurately capture their combined effects on price. Furthermore, the used car market's dynamic nature, influenced by changing consumer preferences and economic conditions, adds another layer of complexity to the prediction task (Lai, 2023).

To address these challenges, this study employs a comprehensive modeling approach that balances traditional statistical methods with modern machine learning techniques. Our methodology includes Ordinary Least Squares (OLS) regression and Elastic Net as linear models that provide clear interpretability of feature relationships, while more sophisticated algorithms—Random Forest and XGBoost—are used to capture complex, non-linear patterns. While linear models offer transparent insights into feature importance and market dynamics, their simplicity may limit their

ability to capture complex patterns. In contrast, the advanced machine learning models often achieve superior predictive accuracy through modeling intricate feature interactions, though at the cost of reduced transparency in their decision-making processes.

Through this comparative analysis, we aim to identify the most effective approaches for used car price prediction while providing practical insights for industry applications. The findings from this study can serve as a foundation for developing more sophisticated pricing tools and informing strategic decisions in the automotive market.

2 Method

This study implements multiple predictive modeling approaches to estimate used car prices. Our analysis employs four distinct models: Ordinary Least Squares (OLS) regression, ElasticNet regularization, Random Forest (RF), and XGBoost. The theoretical foundations and mathematical formulations of these methods are based on Hastie (2009), while the implementation follows the computational framework outlined in Boehmke (2020). The following sections detail each model's methodology, highlighting their mathematical foundations and specific implementations.

2.1 Ordinary Least Squares (OLS) Regression

OLS regression provides our foundational approach to modeling car prices (Saurabh Kumar, 2024). By minimizing the sum of squared residuals, it establishes a linear relationship between features and price under assumptions of independence and consistent variance. The model takes the form:

$$y = \beta_0 + \sum_{i=1}^{p} \beta_i x_i + \epsilon,$$

where y is the target variable, x_i are the predictors, β_i are the coefficients, and ϵ is the error term.

This method offers clear interpretability through its coefficients but faces limitations with non-linear relationships and correlated features. We use it primarily as a baseline model to understand core relationships in our data.

2.2 ElasticNet Regularization

ElasticNet extends OLS by introducing regularization, addressing its limitations in handling high-dimensional and collinear data. It combines ℓ_1-regularization (LASSO) and ℓ_2-regularization (Ridge) to improve predictive performance and feature selection. The objective function for ElasticNet is:

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \left(\alpha \|\beta\|_1 + (1-\alpha) \|\beta\|_2^2 \right) \right\},$$

where λ controls the regularization strength, and α balances the trade-off between ℓ_1 and ℓ_2-regularization.

ElasticNet is particularly beneficial in scenarios where predictors exhibit multicollinearity, as ℓ_2-regularization mitigates the instability of coefficient estimates. Additionally, it is effective in feature selection tasks, with ℓ_1-regularization shrinking irrelevant coefficients to zero. It proves particularly useful for our dataset where various car features may be correlated.

2.3 Random Forest (RF)

Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions to improve accuracy and reduce overfitting. For regression tasks, the prediction is the average of the outputs from individual trees:

$$\hat{y} = \frac{1}{T} \sum_{t=1}^{T} f_t(x),$$

where T is the number of trees and $f_t(x)$ is the prediction from the t-th tree.

The foundation of RF lies in bootstrap aggregating (bagging), where each tree is trained on a random subset of the data, improving generalization by reducing variance. Random Forest excels in capturing complex non-linear relationships and interactions, making it highly effective for predictive tasks. Furthermore, it provides measures of feature importance, offering insights into model behavior, though its interpretability remains limited compared to linear models.

2.4 XGBoost

XGBoost (Extreme Gradient Boosting) is a scalable and efficient implementation of gradient boosting, which builds an ensemble of weak learners (decision trees) to minimize the prediction error iteratively. The objective function combines a loss term and a regularization term:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{t=1}^{T} \Omega(f_t),$$

where l is the loss function, and Ω is a regularization term that penalizes complexity.

XGBoost introduces several innovations, including depth-first tree pruning, column subsampling, and regularization via ℓ_1 and ℓ_2 penalties, all of which enhance generalization and reduce overfitting. XGBoost is well-suited for tasks requiring highly accurate predictions, offering a powerful tool for capturing complex non-linear patterns, though its interpretability is more limited compared to simpler models.

2.5 Evaluation Metrics

The models are evaluated based on their predictive performance on the test dataset. The primary metrics used include Root Mean Squared Error (RMSE), which measures the average magnitude of prediction errors while penalizing larger errors, and Mean Absolute Error (MAE), which represents the mean absolute difference between predictions and actual values. Additionally, the Coefficient of Determination (R^2) quantifies the proportion of variance in the target variable explained by the model. These metrics provide a comprehensive assessment of each model's accuracy and suitability for the task.

2.6 Bootstrapping

Bootstrapping is a statistical technique to check how reliable and accurate a model is by sampling data repeatedly with replacement. In this method, multiple subsets of the original dataset are created through random sampling, where each data point may appear multiple times or not at all. By using these new datasets, bootstrapping helps us see how much the model's predictions can change and how stable they are. This approach is especially useful for small or noisy datasets, as it enables the calculation of confidence intervals, standard errors, and model performance metrics without needing strict rules about the data. Overall, bootstrapping helps better understand the model's abilities and make it more reliable, reducing the chances of overfitting.

3 Data Description and Exploratory Analysis

For this project, we used a dataset from Kaggle containing information about used cars. The dataset includes approximately 188,000 rows, and key features include brand, model year, mileage, fuel type, engine, and accident history. The target variable is the price of the cars.

id	brand	model	model_year	milage	fuel_type	engine	transmission	ext_col	int_col	accident	clean_title	price
0	MINI	Cooper S Base	2007	213000	Gasoline	172.0HP 1.6L 4 Cylinder Engine Gasoline Fuel	A/T	Yellow	Gray	None reported	Yes	4200
1	Lincoln	LS V8	2002	143250	Gasoline	252.0HP 3.9L 8 Cylinder Engine Gasoline Fuel	A/T	Silver	Beige	At least 1 accident or damage reported	Yes	4999
2	Chevrolet	Silverado 2500 LT	2002	136731	E85 Flex Fuel	320.0HP 5.3L 8 Cylinder Engine Flex Fuel Capability	A/T	Blue	Gray	None reported	Yes	13900
3	Genesis	G90 5.0 Ultimate	2017	19500	Gasoline	420.0HP 5.0L 8 Cylinder Engine Gasoline Fuel	Transmission w/Dual Shift Mode	Black	Black	None reported	Yes	45000
4	Mercedes-Benz	Metris Base	2021	7388	Gasoline	208.0HP 2.0L 4 Cylinder Engine Gasoline Fuel	7-Speed A/T	Black	Beige	None reported	Yes	97500
5	Audi	A6 2.0T Sport	2018	40950	Gasoline	252.0HP 2.0L 4 Cylinder Engine Gasoline Fuel	A/T	White	-	None reported	Yes	29950
6	Audi	A8 L 3.0T	2016	62200	Gasoline	333.0HP 3.0L V6 Cylinder Engine Gasoline Fuel	8-Speed A/T	Black	Black	None reported	Yes	28500
7	Chevrolet	Silverado 1500 LTZ	2016	102604	E85 Flex Fuel	355.0HP 5.3L 8 Cylinder Engine Flex Fuel Capability	A/T	White	Gray	None reported	Yes	12500
8	Ford	F-150 XLT	2020	38152	Gasoline	2.7L V6 24V PDI DOHC Twin Turbo	10-Speed Automatic	Snowflake White Pearl Metallic	Black	None reported	Yes	62890

Figure 1: Raw data

We conducted Exploratory Data Analysis (EDA) to understand the dataset, identify potential issues, and explore relationships between features and the target variable, price.

3.1 Missing Values

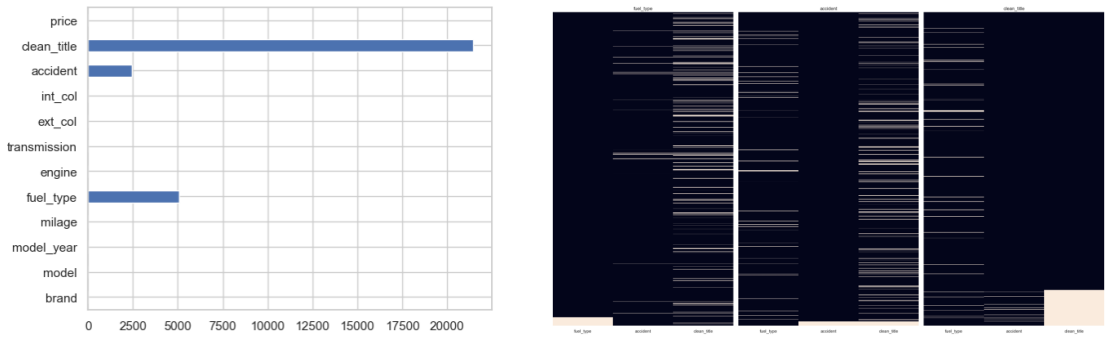

Figure 2: Variables with missing values (left) and heatmaps of the sorted missing values (right)

We first checked for missing values in the dataset. Variables such as fuel type, accident, and clean title had missing data, with clean title having the highest proportion of missing entries. However, we found no significant patterns or correlations between the missing values and other variables.

3.2 Numerical Variables

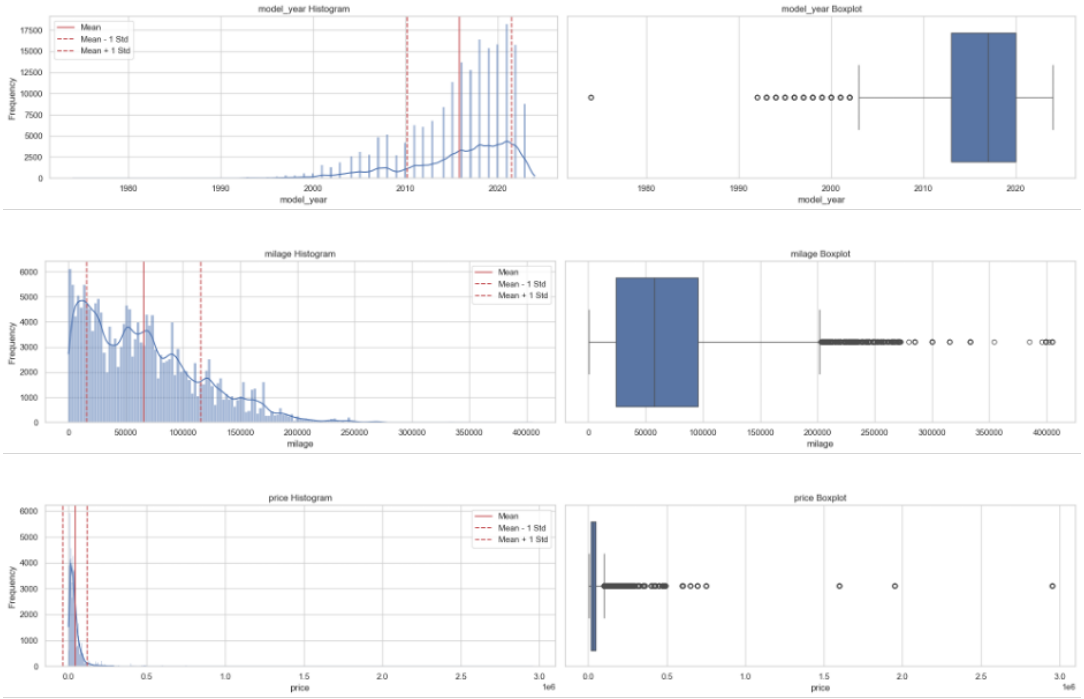

Figure 3: Distribution of numerical and target features

Next, we analyzed the distributions of numerical variables, including model year, mileage, and price, using histograms and boxplots. Most variables showed skewed distributions. For example, price exhibited a strong right skew with a long tail, indicating a significant number of high-value outliers. Model year was concentrated around recent vehicles, with fewer older models present, whereas mileage was heavily distributed toward lower values.

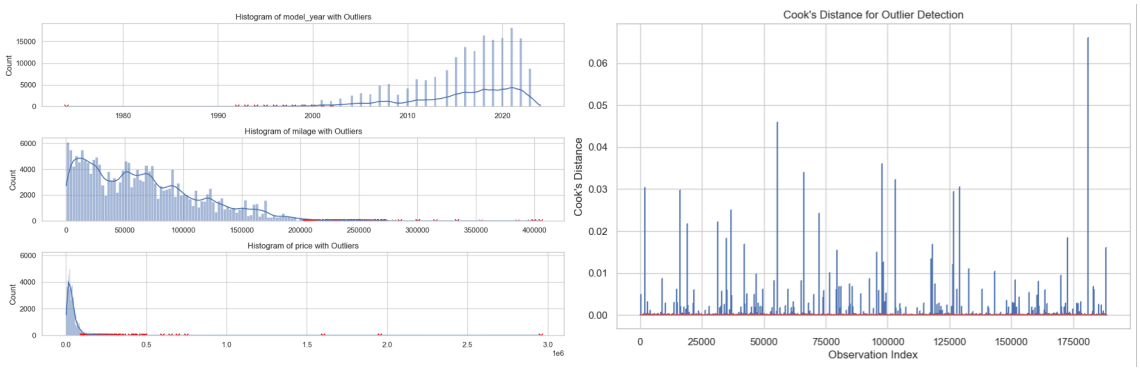

Figure 4: IQR (left) and Cook's Distance (right) for outlier detection

Through outlier analysis using the interquartile range (IQR) and Cook's Distance, we found numerous extreme values, particularly among luxury vehicles with unusually high prices or specifications. In total, 17,326 rows were flagged as outliers using IQR, and 2,846 rows using Cook's Distance. These outliers were removed to ensure model stability, as they could have disproportionately influenced predictions.

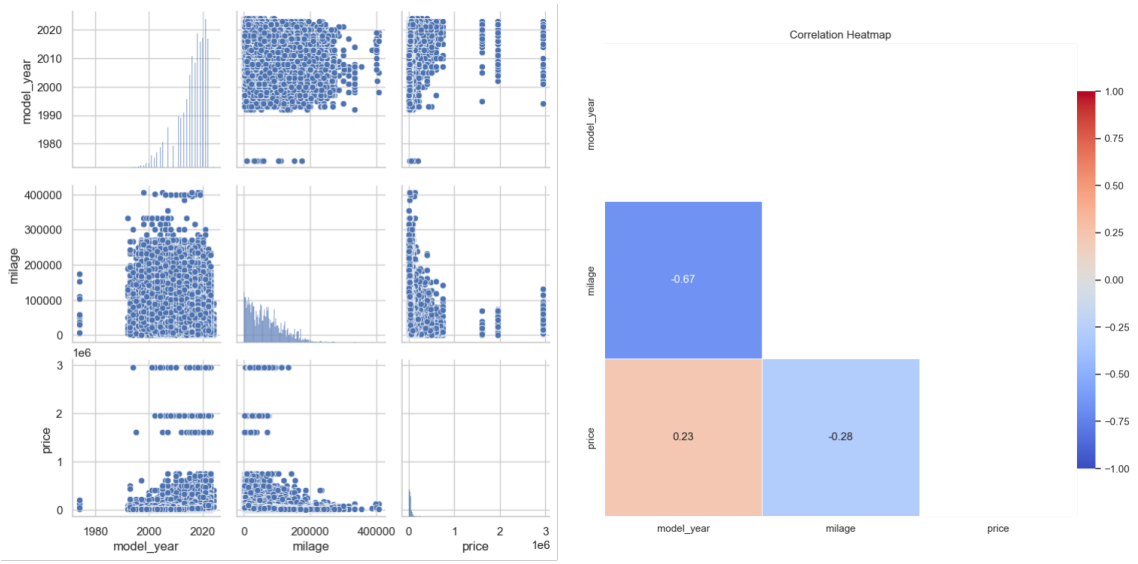

Figure 5: Pairplot (left) and correlation heatmap (right) for numerical variables

Lastly, we examined the relationships between numerical variables and the target variable using correlation heatmaps and pairplots. While correlations were explored, no significant additional insights were found beyond the expected trends.

3.3 *Categorical Variables*

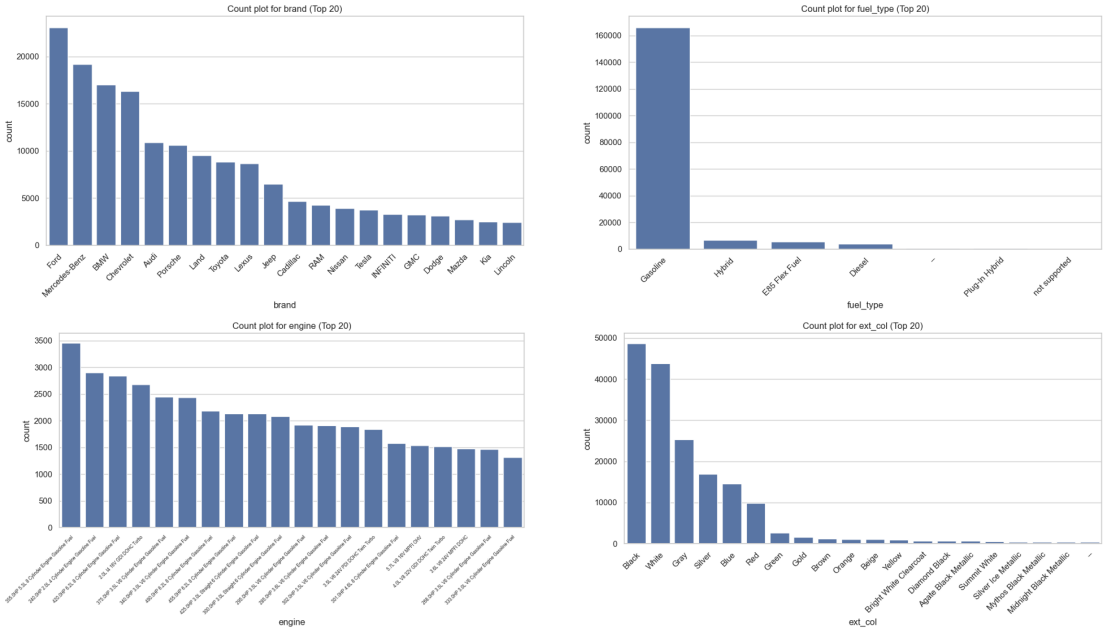

Figure 6: Frequency distributions by categorical features

We also performed EDA on categorical variables, including brand, fuel type, engine, color, and transmission. We used bar plots, boxplots, and violin plots to see these categories' frequency distributions and also how they relate to the target variable. We found substantial price variations across different categories. Luxury brands exhibited significantly higher median prices and wider price ranges compared to economic brands. Additionally, fuel type and engine categories showed notable price differences, with hybrid and V8 engine vehicles commanding premium valuations. The accident history variable proved particularly insightful, with vehicles having no reported accidents commanding higher average prices. Color categories, while less impactful, still demonstrated subtle price variations, with certain colors associated with slightly higher valuations.

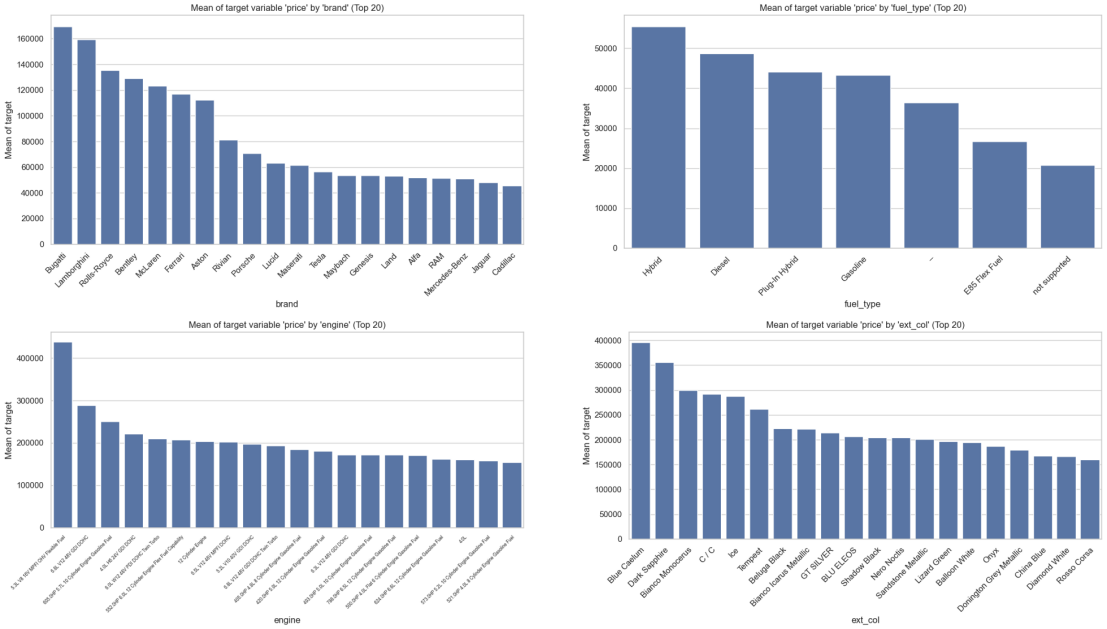

Figure 7: Mean car prices by categorical features

4 Data Preprocessing and Feature Engineering

4.1 Data Preprocessing

The dataset underwent systematic preprocessing to ensure data quality and enhance predictive power. Vehicle brands were categorized into four market segments: Economic, Luxury, Sport/-Luxury, and Off-road/Utility. Technical specifications were standardized by grouping fuel types into five categories (Gasoline, Hybrid, Diesel, Flex Fuel, Unknown), transmission types into five classes (Automatic, Manual, CVT, Auto-Shift, Other), and accident history into three states (No Accident, Accident Reported, Unknown).

	id	model_year	milage	clean_title	price	car_grade	fuel_category	transmission_category	accident_status	hp	displacement	engine_type	color_category	int_color_category
1	0	2007	213000	Yes	4200	Other	Gasoline	Automatic	No Accident	172.0000	1.600000	I4	Other	Gray
2	1	2002	143250	Yes	4999	Luxury	Gasoline	Automatic	Accident Reported	252.0000	3.900000	Other	Gray	Beige
3	2	2002	136731	Yes	13900	Economic	Flex Fuel	Automatic	No Accident	320.0000	5.300000	Other	Other	Gray
4	3	2017	19500	Yes	45000	Luxury	Gasoline	Auto-Shift	No Accident	420.0000	5.000000	Other	Black	Black
5	4	2021	7388	Yes	97500	Luxury	Gasoline	Automatic	No Accident	208.0000	2.000000	I4	Black	Beige
6	5	2018	40950	Yes	29950	Luxury	Gasoline	Automatic	No Accident	252.0000	2.000000	I4	White	Other
7	6	2016	62200	Yes	28500	Luxury	Gasoline	Automatic	No Accident	333.0000	3.000000	V6	Black	Black
8	7	2016	102604	Yes	12500	Economic	Flex Fuel	Automatic	No Accident	355.0000	5.300000	Other	White	Gray
9	8	2020	38352	Yes	62890	Economic	Gasoline	Automatic	No Accident	308.4458	3.402425	V6	White	Black
10	9	2015	74850	Yes	4000	Luxury	Gasoline	Auto-Shift	No Accident	425.0000	3.000000	Other	Black	Other

Figure 8: Cleaned Data

Engine specifications required detailed processing: horsepower (HP) and displacement values were extracted using regular expressions, then categorized by engine type (V6, V8, I4) and displacement range (Small, Medium, Large). Missing values were imputed using within-group means. Color variables were consolidated into major categories to reduce dimensionality. To align with the assumptions of linear regression, the price variable was log-transformed to approximate a normal distribution. This transformation addressed the inherent skewness in the raw price data, improving the model's fit and interpretability.

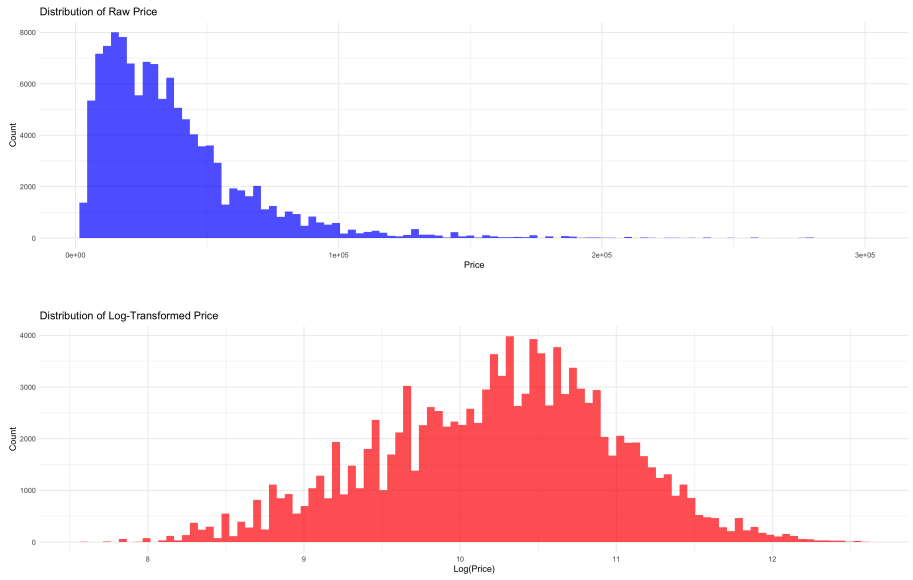

Figure 9: Price Before and After Log-transformation

4.2 Outlier Treatment

Outlier management followed a strict protocol to prevent data leakage. After train-test splitting, outliers were identified in the training set using Cook's Distance with a threshold of 4/n from an initial OLS regression model. Observations exceeding this threshold were removed from the training set only, preserving test set integrity. Approximately 2,300 observations were identified as influential and removed from the dataset. This step was critical in addressing the undue influence of outliers on the model's performance. After removing these observations, the dataset retained sufficient rows for robust analysis, with improved stability and reduced noise.

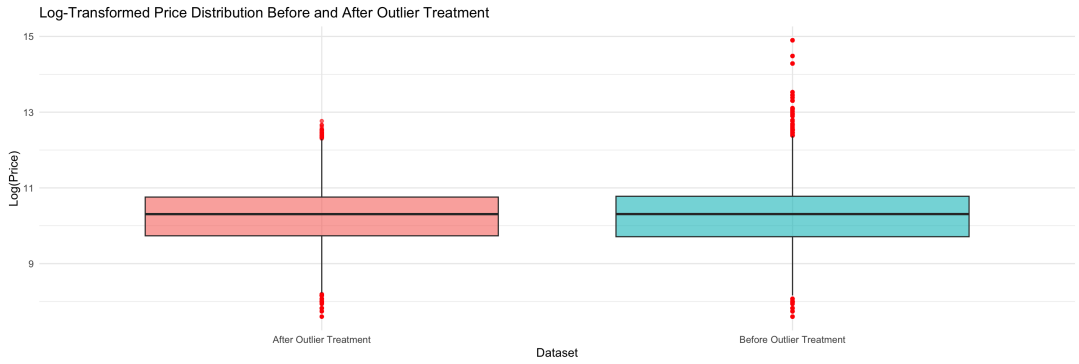

Figure 10: Boxplot of Price Before and After Outlier Treatment

5 Model Evaluation

5.1 OLS Regression Model Results

The results of the OLS regression model predicting the logarithm of car prices are presented. The model includes predictors such as mileage, model year, title condition (clean or not), car grade, fuel type, transmission type, accident history, horsepower, engine displacement, engine type, exterior and interior color categories, and interaction terms between mileage and model year.

For the training data, the model explains approximately 73.13% of the variance in price (R-squared = 73.13%) and exhibits a mean absolute error (MAE) of 10,951.55 and a root mean squared error (RMSE) of 18,722.17. On the test data, the model explains only 11.22% of the variance (R-squared = 11.22%), with an MAE of 16,433.78 and an RMSE of 67,085.40. The discrepancy between the training and test performance suggests potential overfitting or challenges in generalizing to unseen data.

Variable	**Exponentiated Estimate**	**Std. Error**	**t-value**	**p-value**
Mileage	1.00015	7.983×10^{-6}	18.699	< 0.001
Model Year	1.0615	0.0004	138.207	< 0.001
Clean Title (Yes)	1.0989	0.0041	22.890	< 0.001
Car Grade: Luxury	1.0320	0.0028	11.341	< 0.001
Car Grade: Sport/Luxury	1.1670	0.0053	29.361	< 0.001
Fuel Type: Hybrid	0.8156	0.0100	-20.398	< 0.001
Accident Status: No Accident	1.0783	0.0027	27.907	< 0.001
Horsepower	1.0013	1.777×10^{-5}	73.014	< 0.001
Mileage × Model Year	1.0000	3.96×10^{-9}	-19.470	< 0.001

Table I: Exponentiated coefficients for key predictors from the OLS regression model.

While the model captures meaningful relationships between predictors and price, the limited test set performance indicates that further refinements, such as addressing potential overfitting, managing outliers, or improving feature engineering, are necessary to improve the model's generalizability.

5.2 Elastic Net Results

The ElasticNet regression model was applied to predict the logarithm of car prices using a comprehensive set of predictors. A training-test split (80%-20%) was used, with the training data log-transformed for the target variable to stabilize variance. The model underwent hyperparameter tuning using repeated 5-fold cross-validation over a range of α (0 to 1) and λ (0.01 to 0.1) values.

The best-performing model was identified with $\alpha = 0.1$ and $\lambda = 0.01$, selected based on the smallest cross-validated RMSE. During cross-validation, the model achieved an RMSE of 0.5162, an R-squared of 61.45%, and an MAE of 0.3657.

The model's performance was evaluated on the test dataset, with predictions back-transformed to the original price scale using the exponential function. On the test set, the model achieved an RMSE of 75,953.78, an MAE of 30,723.07, and an R-squared of 11.22%. This indicates that the model explains approximately 11.22% of the variance in car prices in the test dataset.

In summary, the ElasticNet model demonstrates moderate predictive capability but has room for improvement in generalizing to new data. Future iterations should focus on refining model complexity and preprocessing steps to enhance performance.

5.3 Random Forest Results

The Random Forest regression model was developed using a dataset with 80% of the data allocated to training and 20% to testing. The hyperparameters were tuned using a 3-fold cross-validation strategy with a grid search. The grid explored two values for the number of predictors randomly sampled at each split (`mtry`), specifically 4 and 6. The minimum node size was fixed at 5, and the variance splitting rule was applied for regression.

The best-performing Random Forest model used an `mtry` value of 4. During cross-validation, the model achieved a RMSE of 69,302.11, an R-squared of 13.09%, and a MAE of 18,251.55. When evaluated on the test set, the model yielded an RMSE of 59,275.15, an R-squared of 15.39%, and an MAE of 17,415.88.

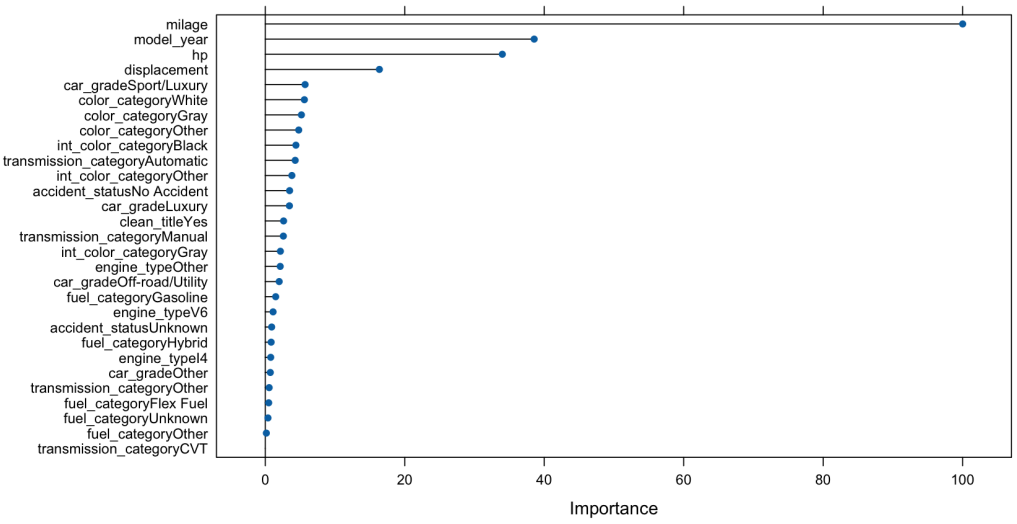

Figure 11: Top Predictors Identified by Random Forest

The results demonstrate that the Random Forest model achieved moderate predictive performance on the test set, with a relatively low R-squared value likely reflecting high variability in car prices and the potential presence of outliers or weakly predictive features. Despite these limitations, the RMSE and MAE suggest that the model provides reasonable predictions for car prices on average. Future work could involve testing additional hyperparameter combinations, such as exploring a larger range of `mtry` values, experimenting with alternative splitting rules, or refining the feature set to enhance model performance.

5.4 XGBoost Results

The XGBoost regression model was developed to predict car prices using a dataset with categorical variables converted to one-hot encoded features. The data was split into training (80%) and testing (20%) sets, with predictors transformed into the matrix format required for XGBoost. The model utilized the gbtree booster with the objective set to minimize squared error. Key hyperparameters included a learning rate (η) of 0.05, a maximum tree depth of 4, a subsampling ratio of 0.8 for training instances, and a subsampling ratio of 0.8 for columns.

To determine the optimal number of boosting rounds, the model underwent 3-fold cross-validation with a maximum of 200 rounds. Based on cross-validation results, the optimal number of boosting rounds was selected, and the final model was trained using this configuration.

The model's performance on the training set showed a RMSE of 68,003.97, a MAE of 17,878.49, and an R-squared of 16.13%. On the test set, the RMSE was 60,006.78, the MAE was 17,393.51, and the R-squared was 16.35%.

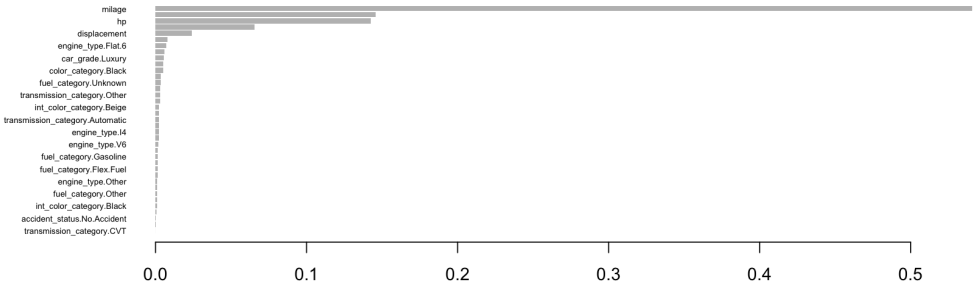

Figure 12: Top Predictors Identified by XGBoost

The results demonstrate that while the XGBoost model provides reasonable predictive accuracy, the relatively low R-squared values suggest that a significant portion of the variance remains unexplained, likely due to inherent complexity or variability in car pricing data. Future work could explore alternative hyperparameter configurations, such as deeper trees or adjusted subsampling rates, as well as additional feature engineering to improve the model's explanatory power.

5.5 Comparison of Model Performance

The table below summarizes the performance of four regression models—OLS (refined), Elastic Net, Random Forest, and XGBoost—based on three key metrics: R-squared, Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

Model	**R-squared**	**RMSE**	**MAE**
OLS	11.22%	67,085.40	16,433.78
Elastic Net	11.22%	75,953.78	30,723.07
Random Forest	15.39%	59,275.15	17,415.88
XGBoost	16.35%	60,006.78	17,393.51

Table II: Performance Comparison of Models

The XGBoost model achieved the highest R-squared value of 16.35%, indicating that it explained the largest proportion of variance in car prices among the four models, closely followed by the Random Forest model with an R-squared of 15.39%. Both ensemble-based models significantly outperformed the Elastic Net and refined OLS models, which both had an R-squared of 11.22%.

In terms of error metrics, Random Forest demonstrated the lowest RMSE of 59,275.15 and the lowest MAE of 17,415.88, indicating the smallest average prediction errors. XGBoost performed comparably, with an RMSE of 60,006.78 and an MAE of 17,393.51. The Elastic Net model exhibited the highest RMSE and MAE values (75,953.78 and 30,723.07, respectively). The OLS model performed better than Elastic Net but was still outperformed by both Random Forest and XGBoost, with an RMSE of 67,085.40 and an MAE of 16,433.78.

Overall, the Random Forest model demonstrated the best balance of explanatory power and predictive accuracy, making it the most effective model for this analysis. The XGBoost model also provided strong performance, highlighting the strength of ensemble-based approaches. These results suggest that ensemble methods such as Random Forest and XGBoost are better suited for predicting car prices in this dataset compared to conventional linear models.

5.6 Bootstrapping

We applied bootstrapping to our final selected model to evaluate its performance stability and prediction accuracy more thoroughly. Using a bootstrapping approach with 1,000 iterations, we resampled the training data and repeatedly trained the model to assess its consistency and reliability. We focused on two main aspects: the model's learning stability and the variability in its predictions. To measure performance stability, we calculated metrics such as the Root Mean Squared Error (RMSE) and its standard deviation.

```
Bootstrap RMSE: 66431.73674460598 ± 19.45903009716405
Bootstrap Mean Prediction: [58747.98474609 68697.07107812  8241.04027734 ... 29686.81164648
 53353.70176563 58173.77899609]
95% Confidence Interval: [[56516.56962891 63192.653125    7973.05380859 ... 27544.10761719
 50828.97158203 54477.21103516], [61202.17089844 74458.92714844  8501.56062012 ... 31705.766162
11
 56085.01132813 62030.36953125]]
```

Figure 13: Bootstrap results

Due to computational constraints in running extensive bootstrapping iterations on our Random-Forest model, the performance metrics are based on preliminary XGBoost bootstrapping results, using a slightly smaller training dataset. Our analysis revealed a mean RMSE of approximately 66,431 with a standard deviation of 19.5, indicating consistent model performance across bootstrap samples. This suggested the model's ability to maintain reliability when trained on different data subsets. While these preliminary results show promise, we anticipate potential improvements in our final model bootstrapping, based on the performance enhancements observed during its development.

We also used bootstrapping to estimate the variability and 95% confidence intervals of the model's predictions to gain insights into the range of potential price estimates. For example, for a specific car, the model might predict a price range between 60,000 and 62,000 units with a 95% probability. These confidence intervals not only validate the model's robustness, but also provide users with valuable information. By understanding the uncertainty in the predictions, users can make more informed decisions, using the model's results to better evaluate potential price ranges.

6 Conclusion

This study developed and evaluated four regression models—OLS, Elastic Net, Random Forest, and XGBoost—to predict used car prices based on features like mileage, model year, horsepower, and others. Below, we summarize the key findings:

6.1 Key Insights from Exploratory Data Analysis

Exploratory Data Analysis (EDA) uncovered important patterns and challenges in the dataset. Missing values were observed in critical variables, including fuel type, accident history, and clean title status. Outlier analysis identified influential cases, such as high-value vehicles and those with extreme mileage, which were addressed to improve model stability. As expected, mileage and model year showed clear trends, with newer, lower-mileage vehicles commanding higher prices. Additionally, categories like luxury brands and hybrid fuel types demonstrated notable price advantages, consistent with market expectations.

6.2 Comparison of Model Performance and Implications for Decision-Making

The OLS regression model highlighted significant coefficients for several predictors. Sport/luxury car grade, clean title, accident history, and model year were key drivers of car prices. Interestingly, hybrid fuel type negatively impacted prices, while newer model years and clean titles positively influenced valuations.

The machine learning models confirmed these findings. In XGBoost, mileage, horsepower, and engine displacement emerged as the most influential features. Similarly, Random Forest identified mileage, model year, and horsepower as the top three predictors. This consistency across models underscores the critical role of these variables in shaping used car prices.

Among the four models, Random Forest struck the best balance between accuracy and interpretability, achieving the lowest RMSE and MAE on the test set. XGBoost also delivered strong predictive performance, though with slightly higher computational demands. In contrast, Elastic Net and OLS regression showed lower predictive accuracy, especially on the test data, underscoring the limitations of linear models in capturing complex relationships.

These findings suggest that ensemble-based methods, such as Random Forest and XGBoost, are better suited for predicting car prices when dealing with high-dimensional and non-linear relationships. These models provide robust and accurate predictions, making them valuable tools for decision-makers involved in pricing, valuation, or inventory management. Future research could incorporate additional features, like market trends or seller conditions, to further enhance the models' explanatory power.

6.3 Limitations and Potential Future Work

The dataset revealed complex, non-linear relationships between features and price, making it difficult to achieve consistently high predictive accuracy across different models. This variability highlights the importance of using models capable of managing such complexity. Models like OLS and Elastic Net, in particular, experienced a notable decline in performance on the test data compared to the training data, indicating challenges in generalizing to unseen data. Addressing overfitting remains a critical area for improvement.

For future work, incorporating external data—such as market trends, regional pricing variations, or seller conditions—could enhance the models' explanatory power and boost predictive accuracy. Additionally, exploring hybrid models or ensemble methods with more extensive hyperparameter tuning may better capture the intricate relationships in car pricing data, leading to greater accuracy and robustness.

References

Boehmke, Greenwell (2020). *Hands-On Machine Learning with R*, 1st ed. Chapman & Hall, p. 484. ISBN: 9781138495685.

Gutierrez, Chris (2024). "3 Ways Auto Dealerships Can Leverage Predictive Analytics", Forbes Agency Council. **available at**: `https://www.forbes.com/councils/forbesagencycouncil/2024/11/22/3-ways-auto-dealerships-can-leverage-predictive-analytics/`.

Hastie Tibshirani, Friedman (2009). *The Elements of Statistical Learning*, 2nd ed. Springer, p. 745. ISBN: 0387848576.

Lai, SYL (2023). *Second-hand car price monitor system*, UTAR Institutional Repository.

Narayana, Chejarla Venkat *et al.*, (2021). "Machine Learning Techniques To Predict The Price Of Used Cars: Predictive Analytics in Retail Business", *2021 Second International Conference on Electronics and Sustainable Communication Systems (ICESC)*, pp. 1680–1687. DOI: `10.1109/ICESC51422.2021.9532845`.

Saurabh Kumar, Avinash Sinha (2024). "Predicting Used Car Prices with Regression Techniques", *International Journal of Computer Trends and Technology*, Vol. 72 No. 6, pp. 132–141. DOI: `10.14445/22312803/IJCTT-V72I6P118`.