Technical Report: Final Project DS 5110: Introduction to Data Management and Processing

Team Members: Zubin Zhang, Zidao Wang, Zhiheng Feng
Khoury College of Computer Sciences
Data Science Program
zhang.zub@northeastern.edu

December 10, 2024

Contents

1 Introduction

The Car Price Prediction Project focuses on developing machine learning models capable of accurately estimating used car prices based on diverse attributes such as manufacturer, model, year, odometer reading, and more. Used car pricing is a multifaceted task, shaped by several factors including market trends, vehicle condition, regional demand, and fuel type. Predicting car prices is not only crucial for buyers and sellers but also beneficial for dealerships, financial institutions, and insurers who rely on accurate price estimations for decision-making.

This project addresses the need for a systematic and data-driven approach to car price prediction. By utilizing advanced machine learning techniques, the project not only improves prediction accuracy but also provides actionable insights into the key factors influencing car prices. Moreover, the inclusion of an interactive Shiny application makes this project practical and user-friendly, enabling real-time predictions for diverse stakeholders.

The primary objectives of the project are threefold, first is to identify and analyze the key factors affecting car prices. Second, to develop predictive models that balance accuracy and interpretability, and third is to provide an intuitive platform for users to interact with the models and receive real-time predictions. The project leverages a comprehensive dataset from a major car listing platform in the United States, ensuring that insights are both robust and relevant.

2 Literature Review

Car price prediction has been extensively studied due to its importance in the automotive industry and consumer decision-making. Machine learning techniques, including Linear Regression, Random Forest, and Gradient Boosting, are commonly used in these studies. Random Forest models excel at capturing non-linear relationships, while Gradient Boosting methods like XGBoost enhance accuracy through iterative error minimization. However, existing research often focuses on specific datasets or regions, making scalability and real-time usability a challenge.

The used car market has seen significant changes, driven by external factors such as the global semiconductor shortage, rising inflation, and growing demand for electric vehicles (EVs). For instance, over 11 million vehicles were removed from global production in 2021 due to the chip shortage, causing a surge in used car prices (Cox Automotive, 2023). The average price of used cars rose to $26,700 in 2021, reflecting the increased demand as buyers sought affordable alternatives to new vehicles (Statista, 2023). Used EVs saw an even sharper rise in value, with prices averaging $58,165 in 2022, driven by factors like rising lithium costs and government incentives for EV adoption.

This project addresses these gaps by combining robust preprocessing methods, advanced machine learning models, and an interactive Shiny app. By integrating scalable solutions with real-time prediction capabilities, the project builds upon existing research to offer both high accuracy and practical usability.

3 Methodology

This project follows a systematic methodology to ensure the reliability and applicability of the results. The process includes data collection, preprocessing, and analysis using advanced machine learning techniques. Below are the details of each step.

3.1 Data Collection

The dataset for this project was sourced from a publicly available car listing platform, encompassing over 400,000 records of used cars in the United States. Key attributes included manufacturer, model, year, price, odometer, condition, fuel type, and region. Given the large size of the raw dataset, a subset of cleaned and processed data was used for analysis to maintain computational efficiency. Initial data handling and cleaning tasks were performed using the R programming language to ensure the dataset was ready for further preprocessing and modeling.

3.2 Data Preprocessing

The preprocessing phase was crucial to ensure high-quality and usable data for model training. The following steps were implemented:

- **Missing Data Handling:** Rows with missing values in critical columns, such as manufacturer and model, were removed to maintain the integrity of the dataset.

- **Outlier Removal:** Outliers in numerical attributes like price and odometer readings were identified and removed using the Interquartile Range (IQR) method and Cook's distance.

- **Data Standardization:** To address inconsistencies, model names (e.g., ”F150” vs. ”F-150”) and manufacturer names were standardized for uniformity.

- **Feature Engineering:** An ID column was added to ensure each record in the dataset was unique, simplifying data management and tracking. Categorical variables, including manufacturer and model, were one-hot encoded for compatibility with machine learning algorithms.

- **Dataset Subsetting:** The dataset was filtered to include only popular manufacturers such as Ford, Toyota, and Chevrolet to focus on robust and actionable insights.

3.3 Analysis Techniques

To ensure robust and accurate predictions, the project employed a variety of analytical models:

- **Linear Regression:** Served as a baseline model to establish relationships between variables and evaluate initial predictive accuracy.

- **Backward and Forward Feature Selection:** Applied to optimize the selection of influential features for model training.

- **Random Forest:** Utilized as a robust ensemble learning technique to capture non-linear relationships and interactions between features.
- **XGBoost:** A highly accurate gradient boosting model that was optimized using hyperparameter tuning to improve prediction performance.
- **Stacked Model:** A meta-model was built to combine the predictions of Random Forest and XGBoost, leveraging the strengths of both models for enhanced accuracy and generalizability.

Each model was evaluated using standard performance metrics such as Root Mean Square Error (RMSE), Mean Absolute Percentage Error (MAPE), and R^2 to assess prediction accuracy, error minimization, and variance explained. Additionally, visualizations, including scatter plots and feature importance charts, were used to interpret and communicate the results effectively. This step-by-step methodology ensured that the project not only delivered reliable results but also offered actionable insights and practical applications.

4 Results

The Car Price Prediction Project delivered significant insights into the determinants of used car prices and the effectiveness of different machine learning models. Below is a detailed summary of the findings:

4.1 Visualizations

The visualization phase of the project aimed to extract meaningful patterns and relationships from the data, providing both an intuitive understanding and strong visual evidence to support the findings. Below is a detailed summary of the visualizations and their interpretations:

4.1.1 Price Distribution

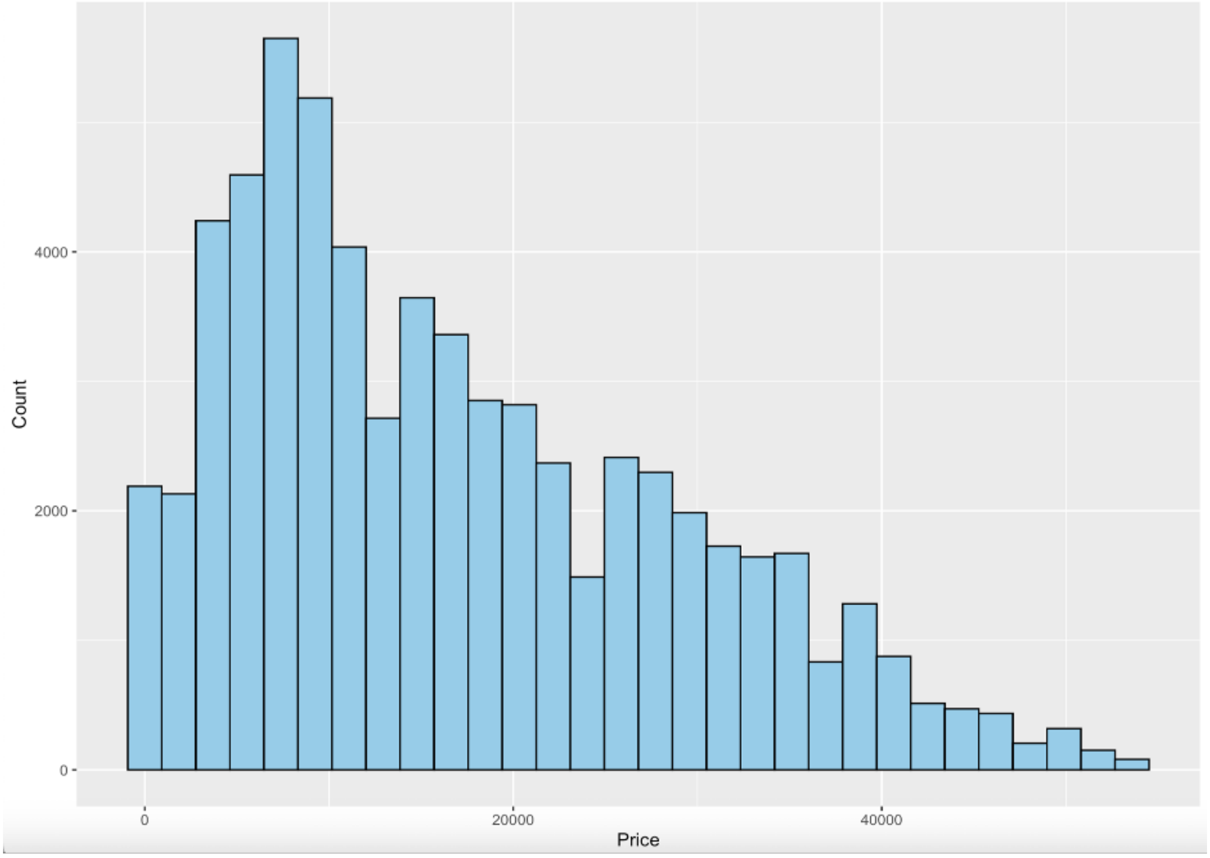


Figure 1: Price Distribution of Vehicles in the Dataset

The bar plot shows the distribution of vehicle prices in the dataset. It highlights a clear skew, with most vehicles priced below $40,000. This distribution indicates the dominance of budget-friendly options in the used car market, though a smaller segment of luxury vehicles is also present. Understanding this distribution is crucial for model development, as it ensures predictions are tailored to the dataset's majority range.

4.1.2 Price vs. Year

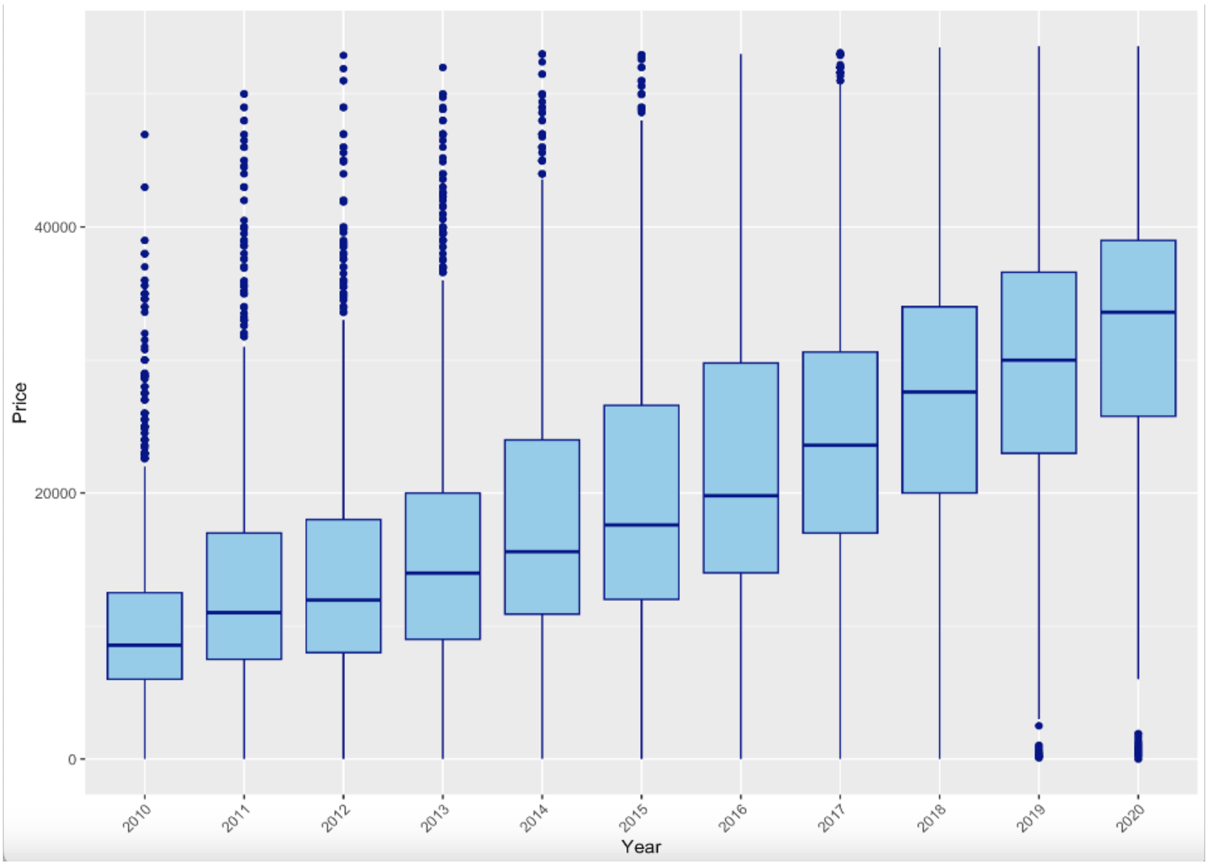


Figure 2: Relationship Between Price and Production Year

The box plot demonstrates the relationship between a car's production year and its price. As expected, vehicles from newer production years have higher median prices, with fewer outliers compared to older models. This trend aligns with market behavior, where newer cars retain higher value due to better features, less wear, and often remaining warranties.

4.1.3 Price vs. Odometer

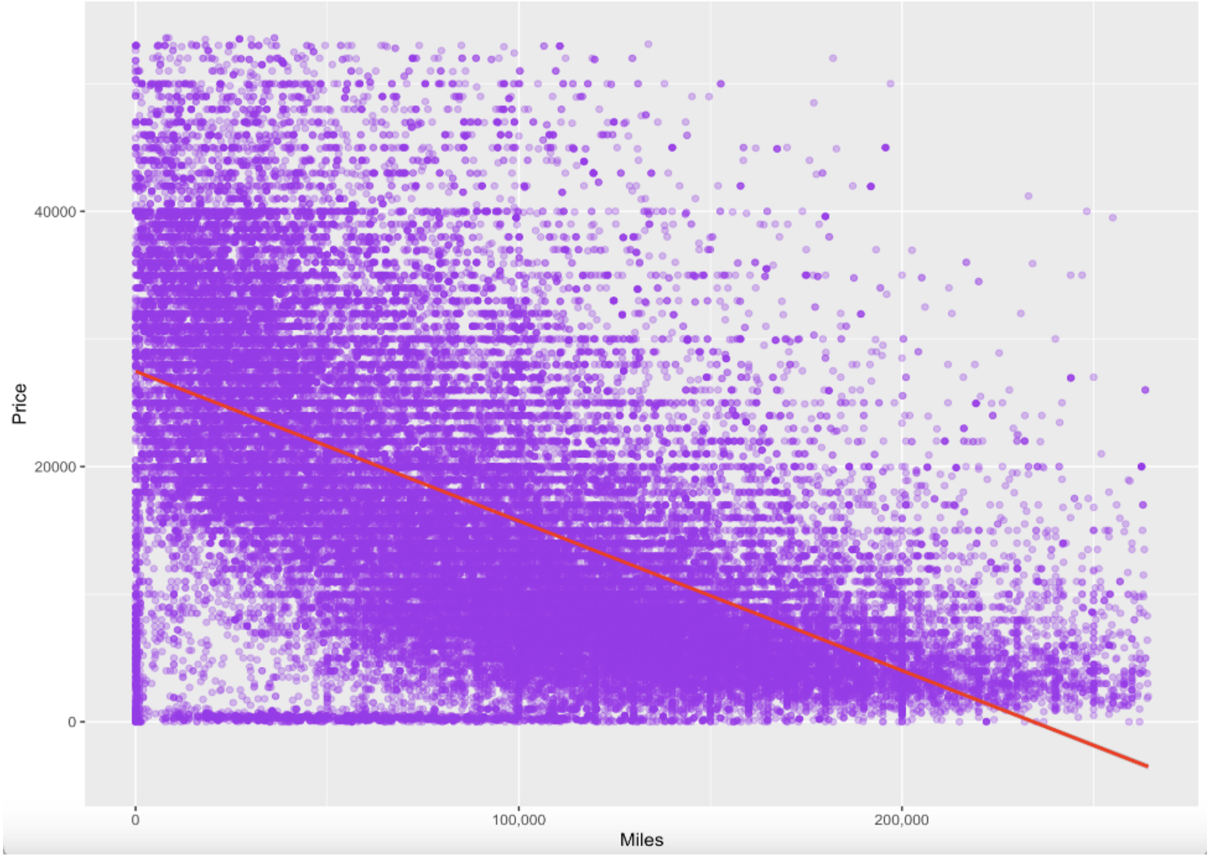


Figure 3: Relationship Between Price and Odometer Reading

The scatter plot illustrates a strong negative correlation between a car's mileage and its price. Cars with higher mileage tend to be significantly cheaper, emphasizing the importance of odometer readings in price predictions. This trend is consistent across manufacturers and model types, reflecting wear-and-tear depreciation.

4.1.4 Manufacturer Distribution

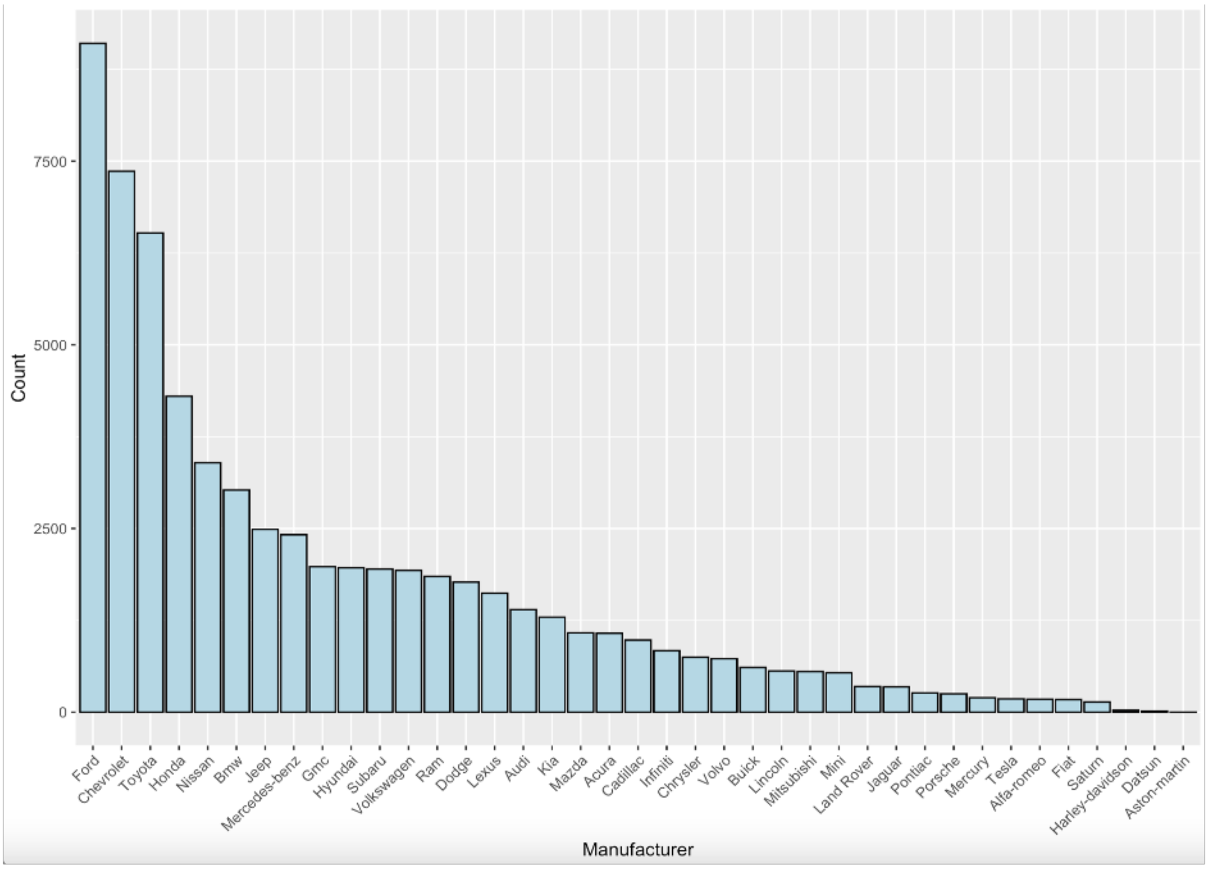


Figure 4: Distribution of Vehicles by Manufacturer

The bar plot shows the number of vehicles listed by each manufacturer. Ford, Chevrolet, and Toyota dominate the dataset, suggesting their popularity in the used car market. This insight informs the focus areas for model feature engineering and analysis, as these manufacturers significantly impact the dataset.

4.1.5 Price vs. Paint Color

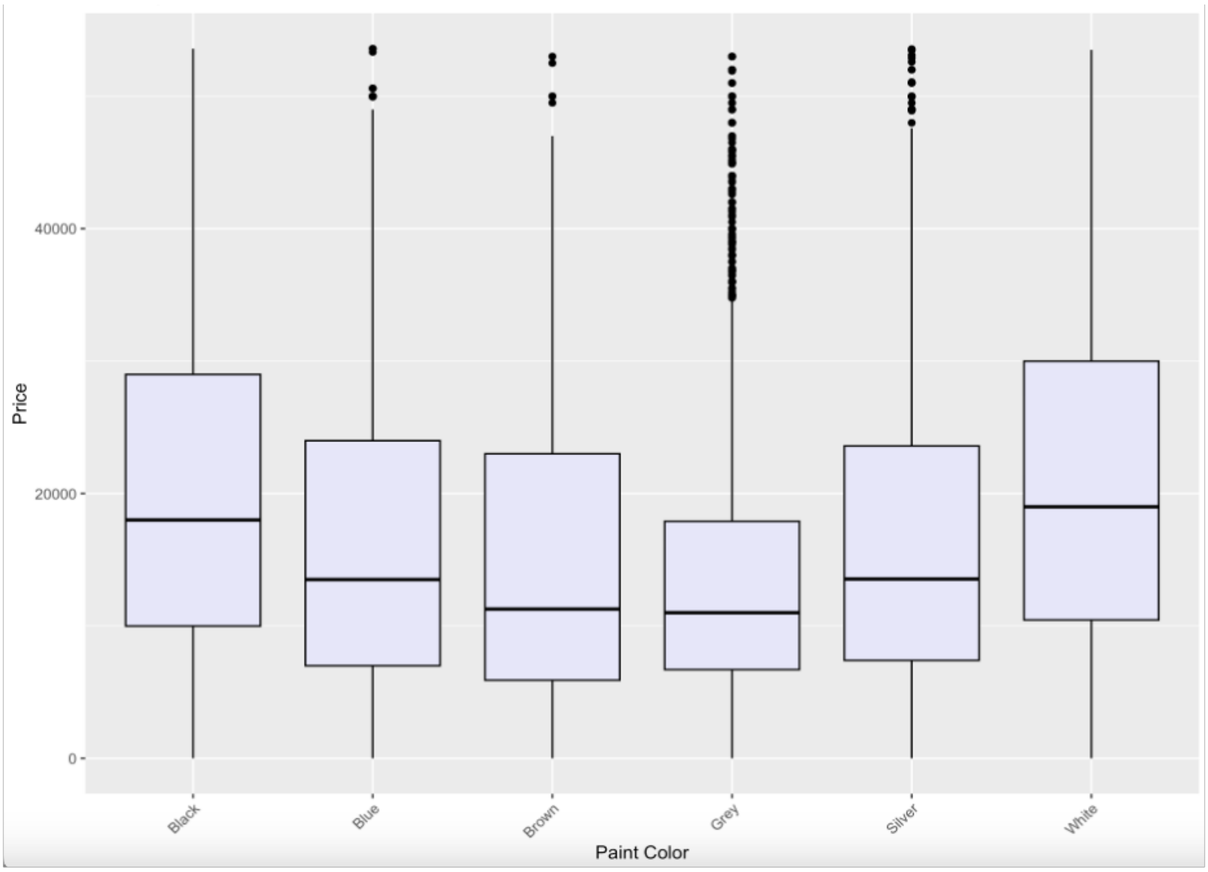


Figure 5: Price vs. Paint Color

The box plot explores the influence of paint color on vehicle pricing. While variations are present, colors such as black, white, and gray are associated with higher median prices. This trend reflects consumer preferences for neutral, widely acceptable colors, which often have better resale value.

4.1.6 Vehicle Age Distribution

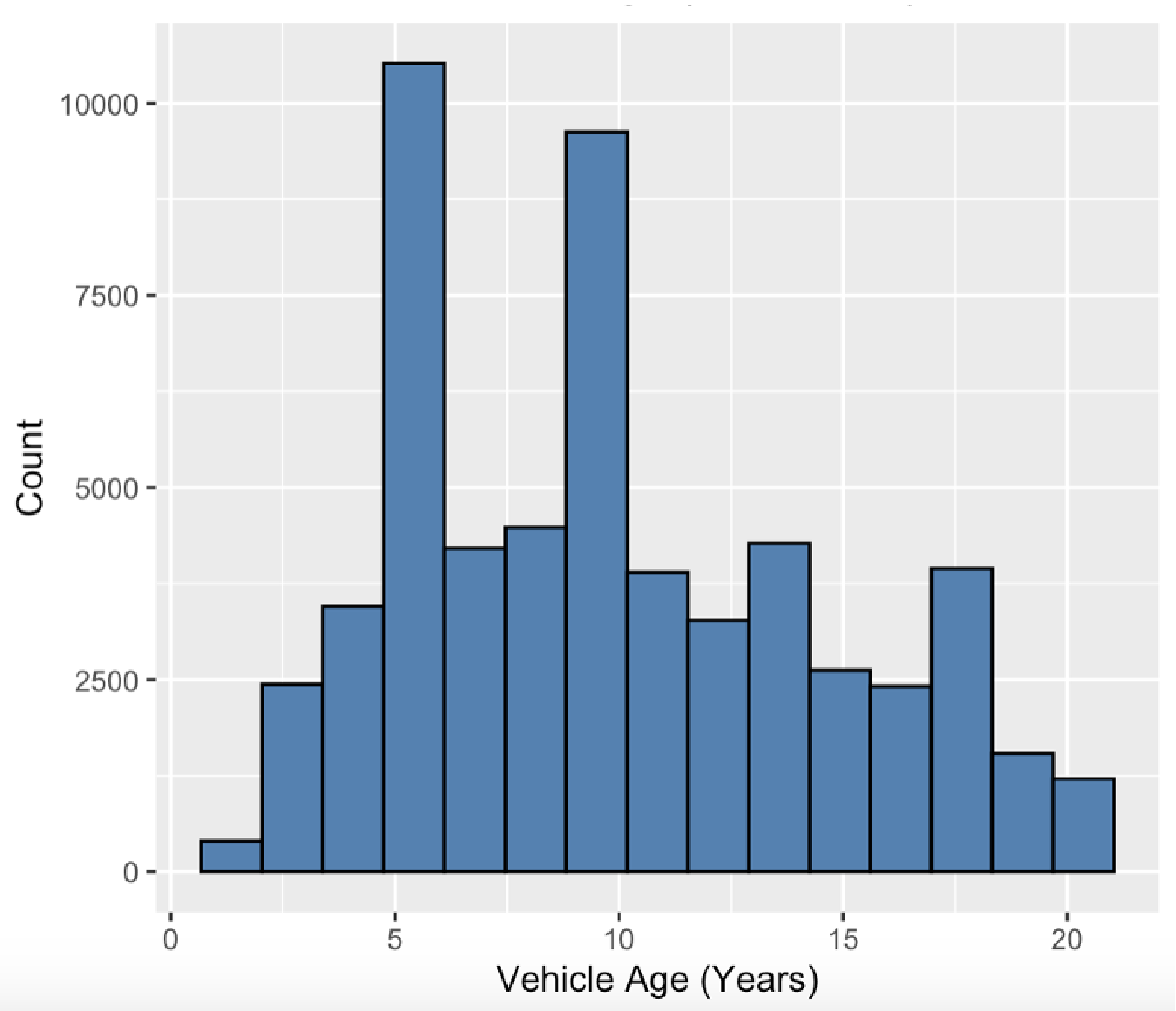


Figure 6: Distribution of Vehicle Ages

The bar plot displays the distribution of vehicle ages in the dataset, showing peaks around specific age ranges. Most vehicles fall within the 5- to 10-year range, indicating a market trend where mid-life vehicles are frequently traded. This insight is vital for understanding the dataset's core focus and ensuring model predictions align with the dominant vehicle age group.

These visualizations provided intuitive insights into the dataset and complemented the model evaluations.

4.2 Linear Regression

Linear regression was utilized as the baseline model to understand the relationships between car prices and their features. Multiple models were developed to evaluate different approaches to feature selection and their impact on prediction accuracy.

4.2.1 Model Overview

Linear regression models explored various combinations of features, from high-correlation predictors to all available attributes. The analysis aimed to balance simplicity, interpretability, and predictive performance.

4.2.2 High-Correlation Features Model

	Price	Year	Odometer
Price	1.0000000	0.5651999	-0.6260485
Year	0.5651999	1.0000000	-0.6059002
Odometer	-0.6260485	-0.6059002	1.0000000

Table 1: Correlation Matrix of Price, Year, and Odometer.

This model included only the variables **year** and **odometer**, which demonstrated significant correlations with car prices:

- RMSE (Root Mean Square Error): 8,304
- MAPE (Mean Absolute Percentage Error): 10.86%
- R-squared: 44.57%

While the model provided insights into the primary factors influencing car prices, its performance was limited due to the exclusion of other meaningful predictors. The results indicated that relying solely on high-correlation features is insufficient for accurate predictions.

4.2.3 All Features Model

The all-features model incorporated all available attributes, significantly improving predictive performance:

- RMSE: 5,102
- MAPE: 7.57%
- R-squared: 79.07%

By leveraging the full dataset, this model captured additional complexities and interactions among variables, leading to more accurate predictions. However, its interpretability was reduced due to the inclusion of numerous predictors.

4.2.4 Backward Feature Selection

Backward feature selection iteratively removed less significant variables, retaining only the most impactful predictors:

- Selected Features: **id**, **year**, **manufacturer**, **odometer**, **model**

- RMSE: 5,102
- MAPE: 7.57%
- R-squared: 79.07%

4.2.5 Forward Feature Selection

Forward feature selection added predictors step-by-step based on their contribution to the model. The final selected features were identical to those of the backward selection model:

- Selected Features: **id**, **year**, **manufacturer**, **odometer**, **model**
- RMSE: 5,102
- MAPE: 7.57%
- R-squared: 79.07%

Both forward and backward selection highlighted the significance of the same features, emphasizing their importance in predicting car prices effectively.

4.2.6 Cross-Validated Model

To ensure the model's stability and generalizability, a 10-fold cross-validation was performed:

- RMSE: 5,102
- MAPE: 7.57%
- R-squared: 79.07%

The cross-validation results confirmed the consistency of the model's performance across different data splits.

4.2.7 Visualizations

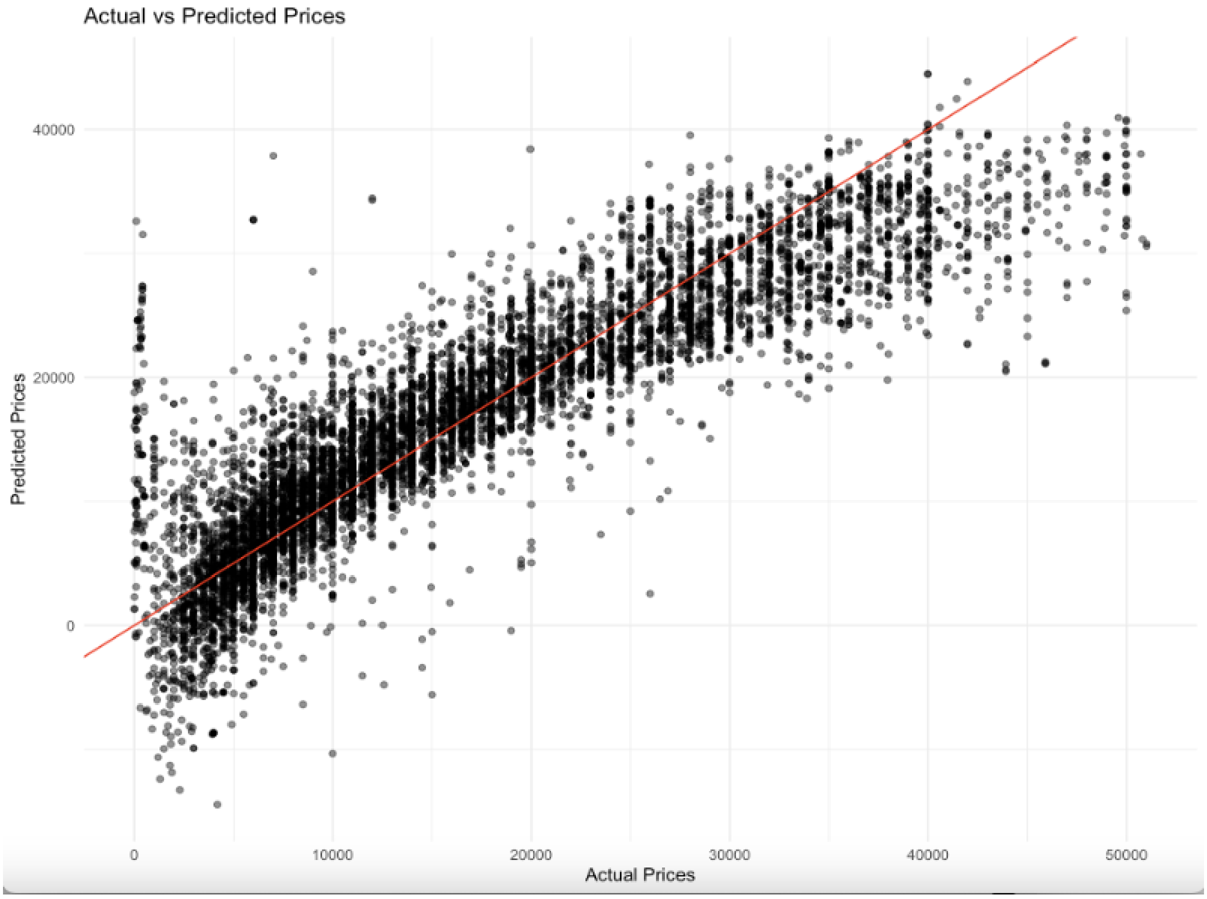


Figure 7: Actual vs. Predicted Prices

The scatter plot shows a strong alignment of predicted prices with actual prices, indicating the model's reliability. Most points cluster around the diagonal, representing accurate predictions.

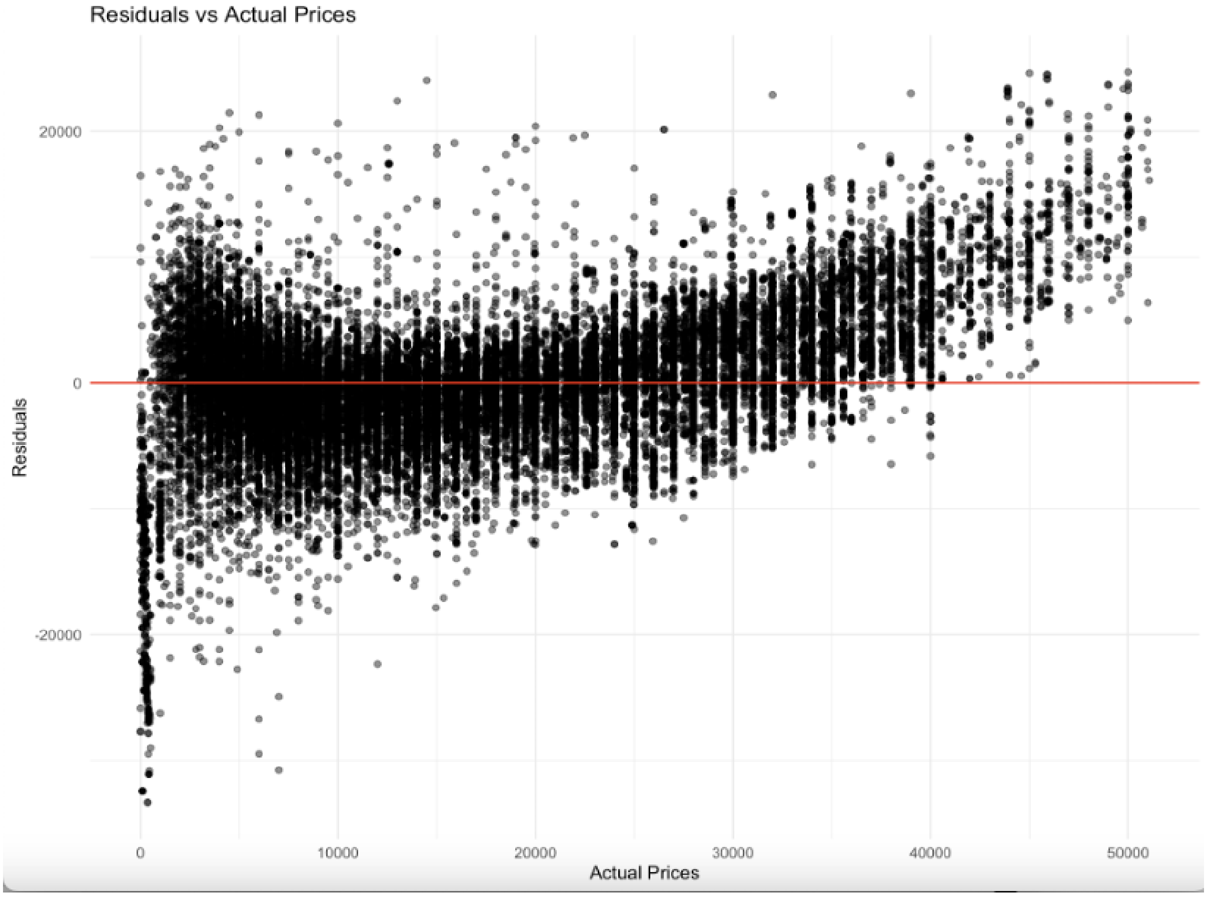


Figure 8: Residuals vs. Actual Prices

The residual plot demonstrates an even distribution of errors around the horizontal zero line, with minor heteroscedasticity observed at higher price ranges. This pattern suggests room for improvement in handling variability at extreme values.

4.2.8 Summary

Linear regression provided a strong baseline for understanding the relationships between car attributes and prices. The inclusion of key predictors like **year**, **manufacturer**, **odometer**, and **model** significantly improved accuracy. However, the model's inability to capture non-linear interactions limits its effectiveness compared to more advanced ensemble methods, which will be explored in subsequent analyses.

4.3 Random Forest

The Random Forest model significantly improved prediction accuracy by capturing non-linear relationships and complex interactions among features. Trained with 500 trees, it demonstrated robust performance:

- **RMSE (Root Mean Square Error):** 4,352.78
- **MAPE (Mean Absolute Percentage Error):** 7.24%
- **R-squared:** 84.77%

4.3.1 Feature Importance

The most influential features identified by the model are as follows:

- **Model:** Highest impact on price.
- **Year:** Newer cars are priced higher.
- **Odometer:** Higher mileage reduces value.
- **Manufacturer:** Reflecting brand reputation.

The feature importance plot above visualizes the contributions of each variable to the model. The **model** variable had the highest importance, indicating its significant role in predicting car prices. In contrast, **ID** had minimal influence, serving primarily as a unique identifier.

4.3.2 Visualization

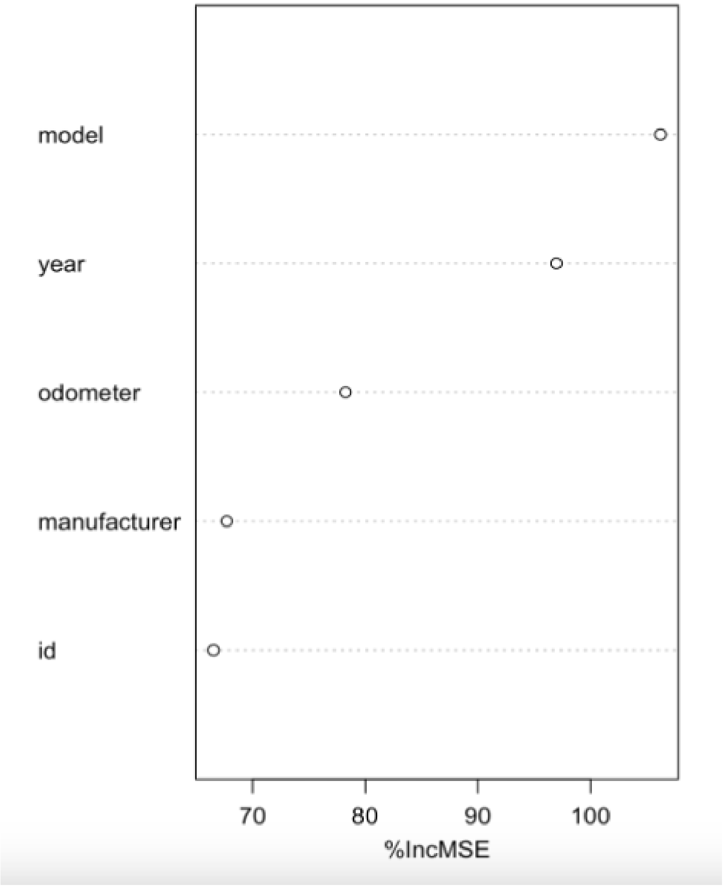


Figure 9: Feature Importance Plot for the Random Forest Model.

This plot clearly illustrates that **model**, **year**, and **odometer** are the most critical predictors of car prices. The %IncMSE metric on the x-axis represents the relative importance of each feature, showing how much each variable contributes to reducing prediction errors. Features with higher values, such as **model**, are essential for accurate predictions, while features like **ID** contribute negligibly.

4.3.3 Summary

The Random Forest model outperformed linear regression, explaining 84.77% of price variance and significantly reducing errors. Its ability to leverage critical predictors like **model** and **year** highlights its suitability for car price prediction tasks.

4.4 XGBoost

The XGBoost model further refined the predictions by leveraging advanced gradient-boosting techniques and iterative optimization. This model outperformed both Linear Regression and Random Forest in terms of efficiency and accuracy:

- **RMSE (Root Mean Square Error):** 4,346.5
- **MAPE (Mean Absolute Percentage Error):** 6.93%
- **R-squared:** 84.82%

4.4.1 Feature Engineering and Preprocessing

To prepare the data for XGBoost, categorical features such as **manufacturer** and **model** were one-hot encoded, converting them into numerical format suitable for gradient boosting algorithms. This preprocessing ensured that the model could effectively utilize all features for prediction.

4.4.2 Performance Insights

The XGBoost model showed slightly better performance than the Random Forest model with a marginally lower RMSE and higher R-squared value. This improvement can be attributed to the model's ability to minimize residual errors using gradient boosting. Additionally, XGBoost efficiently handled the high-dimensional dataset, demonstrating its strength in both computational efficiency and predictive accuracy.

4.4.3 Summary

The XGBoost model proved to be a powerful tool for predicting used car prices, leveraging both advanced optimization techniques and efficient feature utilization. With its strong performance metrics, it slightly edged out Random Forest, showcasing its ability to capture complex interactions and subtle patterns in the dataset. The model's reduced errors and higher explanatory power make it a valuable component of this car price prediction pipeline.

4.5 Stacked Model

The stacked model combined the predictions from the Random Forest and XGBoost models using a linear meta-model, significantly improving predictive accuracy and reducing errors. By leveraging the complementary strengths of both base models, the stacked model achieved the best performance metrics in this project:

- **RMSE (Root Mean Square Error):** 4,091.86
- **MAPE (Mean Absolute Percentage Error):** 6.24%
- **R-squared:** 86.54%

4.5.1 Methodology and Process

The stacked model was developed by combining the predictions from Random Forest and XGBoost into a single dataset. A linear regression model served as the meta-model to learn the relationship between these predictions and the actual target values. This approach allowed the stacked model to aggregate the strengths of both base models, effectively addressing their individual weaknesses.

4.5.2 Performance Insights

The stacked model demonstrated superior performance compared to the individual models, achieving the lowest RMSE and highest R-squared values. This result indicates that combining predictions from multiple models can provide a more accurate and robust estimate than relying on a single model.

The reduced MAPE further highlights the model's ability to minimize relative prediction errors, making it highly suitable for practical applications where precision is critical.

4.5.3 Summary

The stacked model emerged as the most accurate predictive model in this project, showcasing the value of ensemble learning techniques. By combining the feature exploration capabilities of Random Forest with the optimization strengths of XGBoost, the stacked model provided a comprehensive and reliable framework for car price prediction. This approach not only reduced errors but also enhanced the overall predictive power, making it an ideal choice for future extensions and real-world implementations.

4.6 Comparison of Models

The table below summarizes the performance of the models based on RMSE and R-squared, providing a clear comparison of their accuracy and explanatory power:

Model	RMSE	R-squared
Linear Regression (High-Correlation Features)	8304.11	44.57%
Linear Regression (All Features)	5102.96	79.07%
Linear Regression (Feature Selection)	5102.96	79.07%
Random Forest	4352.78	84.77%
XGBoost	4346.50	84.82%
Stacked Model	4091.86	86.54%

Table 2: Model Performance Summary Based on RMSE and R-squared.

4.6.1 Insights from Comparison

- **Linear Regression (High-Correlation Features):** Simplistic and interpretable but with limited performance (RMSE: 8304.11, R-squared: 44.57%).
- **Linear Regression (All Features):** Improved performance by including all variables (RMSE: 5102.96, R-squared: 79.07%).
- **Linear Regression (Feature Selection):** Similar performance to the all-features model due to optimized feature selection (RMSE: 5102.96, R-squared: 79.07%).
- **Random Forest:** Robust and capable of capturing non-linear relationships, significantly outperforming Linear Regression (RMSE: 4352.78, R-squared: 84.77%).
- **XGBoost:** Marginally better than Random Forest, benefiting from advanced optimization techniques (RMSE: 4346.50, R-squared: 84.82%).
- **Stacked Model:** The best performer by integrating Random Forest and XGBoost predictions, achieving the highest accuracy (RMSE: 4091.86, R-squared: 86.54%).

The results highlight the superiority of ensemble techniques, with the Stacked Model offering the most accurate and reliable predictions. This approach leverages the strengths of individual models, making it ideal for this project.

5 Discussion

The results of this project highlight the effectiveness of advanced machine learning models, particularly ensemble methods like Random Forest and XGBoost, in predicting used car prices. Linear Regression, while interpretable, struggled to capture the complex, non-linear relationships inherent in the dataset, as evidenced by its lower R-squared value (44.57%) and higher RMSE (8304.11). Ensemble methods addressed these limitations by combining multiple decision trees and optimizing predictions, leading to significantly better performance. The Stacked Model further enhanced accuracy by leveraging the complementary strengths of Random Forest and XGBoost, achieving the lowest RMSE (4091.86) and the highest R-squared (86.54%).

These findings align with existing literature, which emphasizes the robustness of Random Forest and XGBoost in handling non-linear data and reducing overfitting. However, the project's results surpass those of prior studies by integrating these models into a stacked framework, providing even greater predictive power. One notable observation is the high importance of features such as vehicle model, year, and odometer reading, which were consistent across models and matched prior research highlighting these variables as critical price determinants.

Discrepancies arose in the lack of features like accident history and regional market conditions, which limited the model's ability to capture all price variations. Additionally, while XGBoost showed efficiency in optimization, its performance gain over Random Forest was marginal, suggesting diminishing returns when additional complexity is introduced without more diverse features. The results underscore the need for richer datasets and advanced techniques like stacking to improve predictive accuracy.

6 Conclusion

This project successfully demonstrated the potential of machine learning models to predict used car prices with high accuracy. Key findings include the superior performance of ensemble models, particularly the Stacked Model, which achieved the highest R-squared (86.54%) and lowest RMSE (4091.86). Critical features such as vehicle model, year, and odometer reading were found to have the most significant influence on price predictions.

However, the project has limitations. The dataset lacked features such as accident history, warranty information, and regional market trends, which could further enhance model performance. Additionally, while the models were effective, their interpretability decreased as complexity increased, posing challenges for practical implementation in consumer-facing applications.

Future research should focus on expanding the dataset to include more comprehensive features, exploring deep learning models for even greater accuracy, and evaluating the models' scalability to larger, more diverse datasets. Integrating real-time data streams and user feedback into the predictive framework could also improve usability and relevance, ensuring the models remain adaptable to changing market conditions.

7 References

1. Home. USA - Flash report, Automotive sales volume, 2022 - MarkLines Automotive Industry Portal. (2022). https://www.marklines.com/en/statistics/flash_sales/automotive-sales-in-usa-by-month-2022#:~:text=Total%202022%20full%2Dyear%20sales,to%202%2C979%2C113%20units%20for%202022

2. Carlier, M. (2024, March 19). Topic: Used vehicles in the United States. Statista. https://www.statista.com/topics/9879/used-vehicles-in-the-united-states/#topicOverview

3. Cox Automotive Forecast: U.S. auto sales expected to finish 2023 up more than 11% year over year, as General Motors retains top spot, Hyundai Motor Group jumps past Stellantis. Cox Automotive Inc. (2023, December 27). https://www.coxautoinc.com/news/cox-automotive-forecast-december-2023-u-s-auto-sales-forecas
#:~:text=Cox%20Automotive%20forecasts%20full%2Dyear,of%2017.5%20million%20in%202016

4. Zhang, Z. (2024, October). Zubinzhang1997/used-car-analysis. GitHub. https://github.com/zubinzhang1997/Used-Car-Analysis.git

A Appendix A: Code

The complete code files for this project is extensive, as it includes multiple scripts and configurations for data preprocessing, model training, and visualization. To ensure transparency and reproducibility, all the code files have been uploaded to the project's GitHub repository. You can access them at the following link:

GitHub Repository - Used Car Analysis.

B Appendix B: Additional Figures

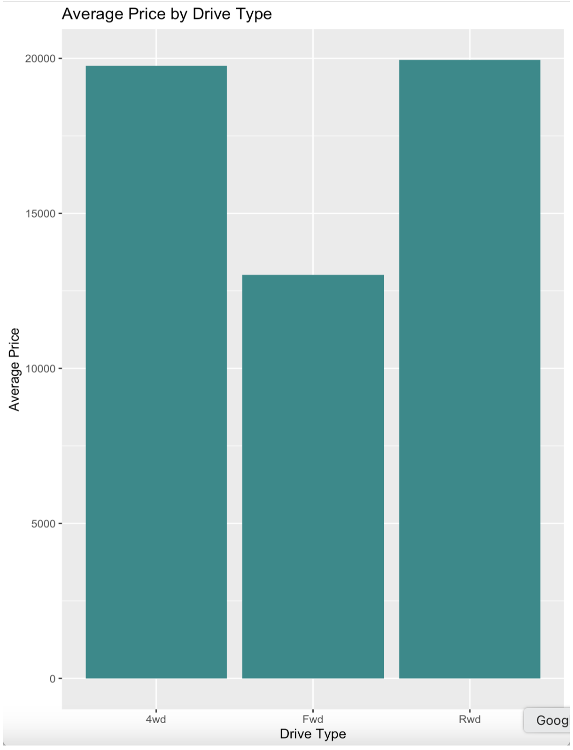


Figure 10: Average Price by Drive Type.

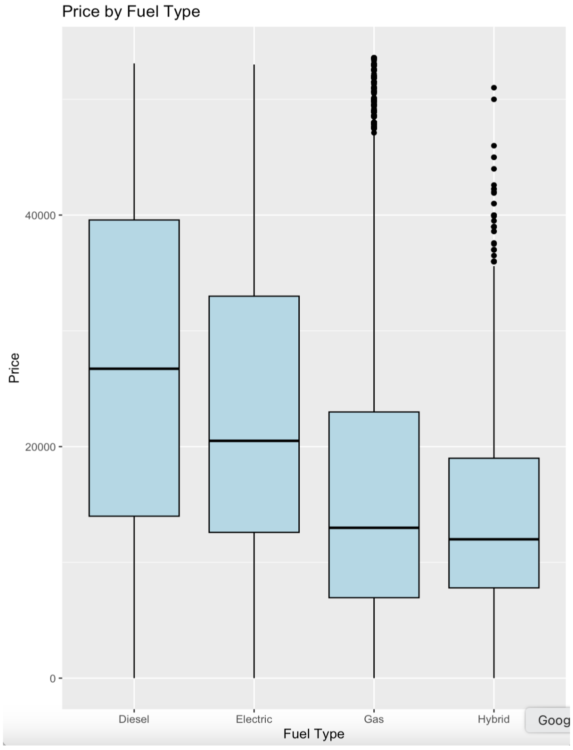


Figure 11: Price by Fuel Type.

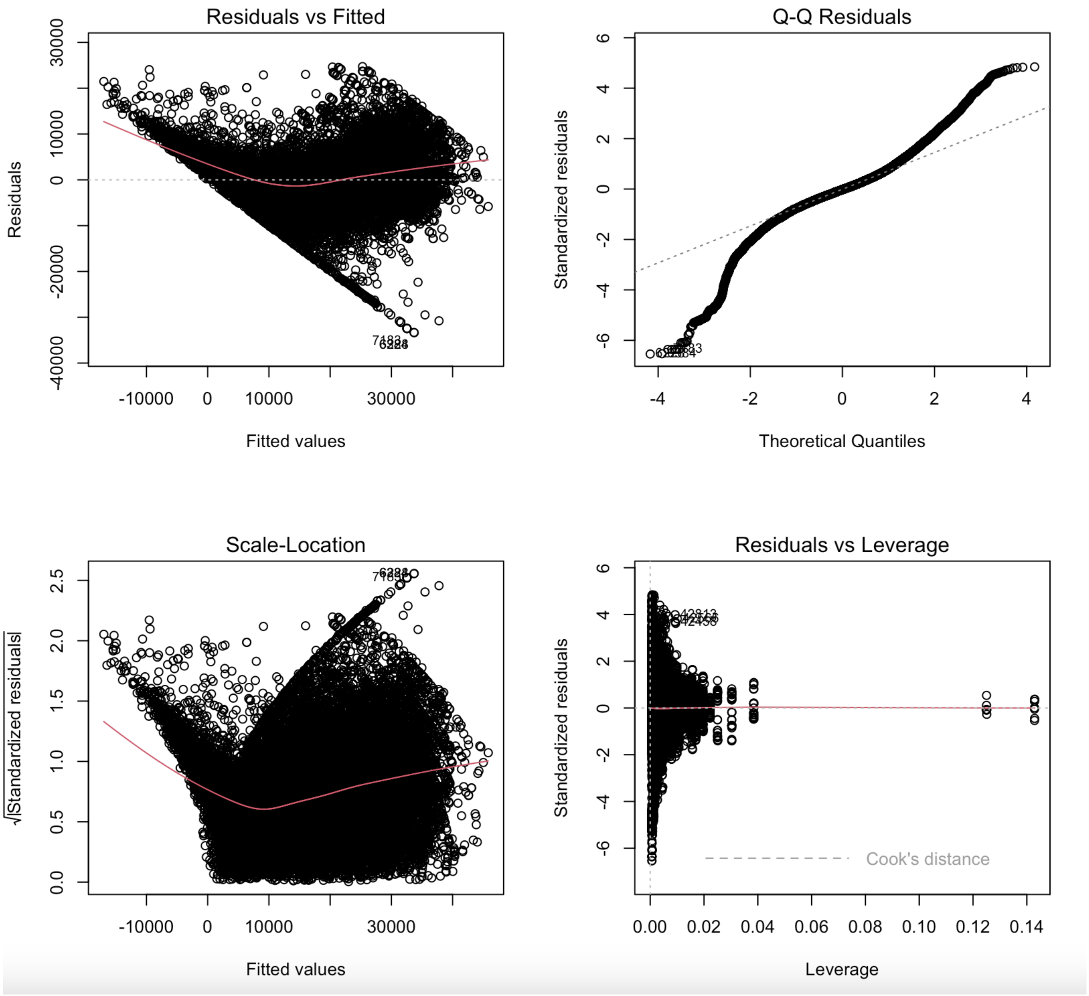
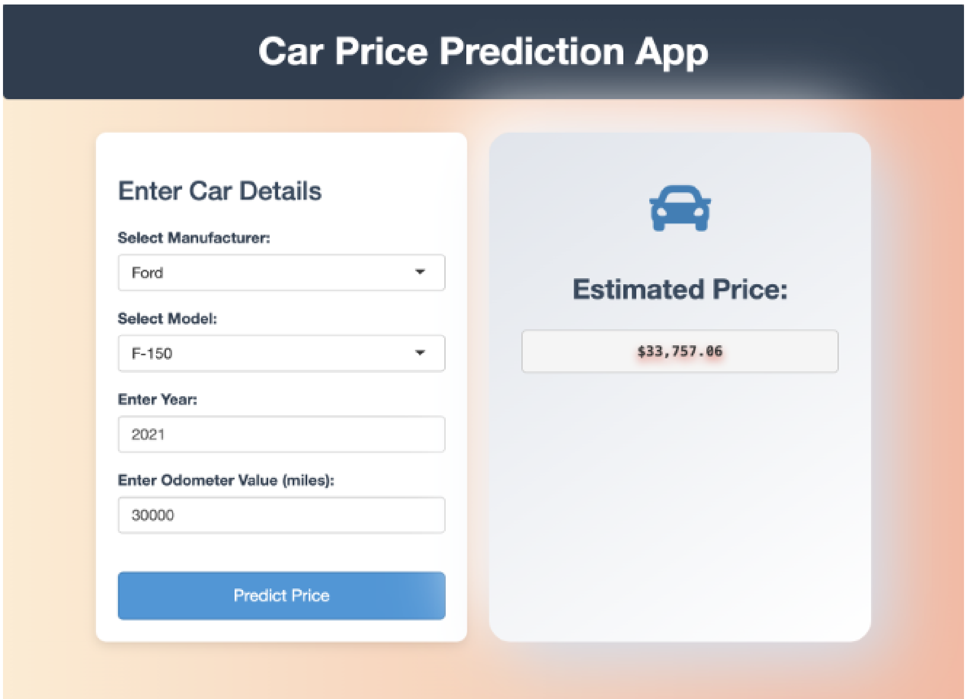


Figure 12: Residual Diagnostics for Linear Regression Model Evaluation.



Car Price Prediction App

Enter Car Details

Select Manufacturer: Ford

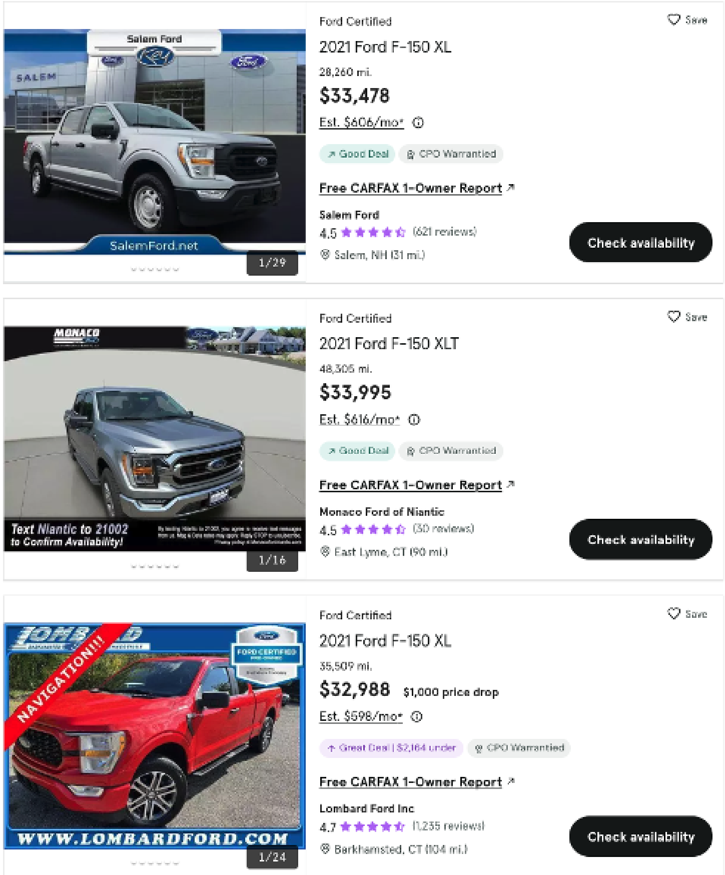
Select Model: F-150

Enter Year: 2021

Enter Odometer Value (miles): 30000

Estimated Price: $33,757.06

Figure 13: Example of Predicted Price Using the Car Price Prediction App.



Ford Certified
2021 Ford F-150 XL
28,260 mi.
$33,478
Est. $606/mo*
Good Deal | CPO Warrantied
Free CARFAX 1-Owner Report
Salem Ford
4.5 ★★★★☆ (621 reviews)
Salem, NH (31 mi.)

Ford Certified
2021 Ford F-150 XLT
48,306 mi.
$33,995
Est. $616/mo*
Good Deal | CPO Warrantied
Free CARFAX 1-Owner Report
Monaco Ford of Niantic
4.5 ★★★★☆ (30 reviews)
East Lyme, CT (90 mi.)

Ford Certified
2021 Ford F-150 XL
35,509 mi.
$32,988 $1,000 price drop
Est. $598/mo*
Great Deal | $2,164 under | CPO Warrantied
Free CARFAX 1-Owner Report
Lombard Ford Inc
4.7 ★★★★☆ (1,235 reviews)
Barkhamsted, CT (104 mi.)

Figure 14: Real-Time Price Results Retrieved from Online Car Listings.