PREDICTING USED CAR VALUE USING STATISTICAL AND MACHINE LEARNING METHODS WITH DATA FROM ONLINE PLATFORM

Te Shi
University at Buffalo
teshi@buffalo.edu

Chao Wu
University at Buffalo
cwu64@buffalo.edu

Jiabao Yao
University at Buffalo
jyao27@buffalo.edu

Shijie Zhou
University at Buffalo
shijiezh@buffalo.edu

ABSTRACT

Used car pricing influences buyer and seller decisions as well as broader economic factors such as employment and market trends. Accurate prediction of resale values requires analyzing key factors like brand, mileage, ownership history, and accident records. Existing studies are limited in datasets and model selection, failing to address the complexity of large, unstructured, and inconsistent data. This study addresses these challenges by collecting comprehensive data from CarMax.com, performing standard exploratory data analysis (EDA) to uncover patterns and correlations. We evaluate multiple statistical and machine learning models, including random forests, XGBoost, and Catboost, diving into experimental deep learning methods, to identify effective approaches for predicting used car prices. The results demonstrate a overview of reliability of various models in terms of interpretability and predictive accuracy, offering insights into factors influencing resale prices and tools to support transparent and data-driven decision-making in the used car market.

***K*eywords** Used Car Prediction · Used Car Pricing · EDA · Statistical Machine Learning · Deep Learning Application · MySQL Database · Streamlit · Data Persistence · Data Product

1 Introduction

Used car pricing is a key issue in the automotive industry because it not only affects the decisions of buyers and sellers, but may also affect other invisible factors such as employment, market trends, stocks, etc. Cars are a necessity in social life, but not everyone can afford a new car. Therefore, studying market laws such as used car pricing also has an impact on people's livelihood. Predicting the resale value of used cars accurately requires understanding the main factors influencing price, such as brand, mileage, number of owners, class, fuel type, damages, and accidents. This study aims to develop data-driven models to predict the resale price of used cars and identify the key attributes that influence pricing.

Previous studies and projects have addressed the issue of used car pricing[1][2][3]. However, these efforts often failed to address the complexity of the problem. Many rely on official datasets such as from Kaggle and these datasets fail to capture real-world scenarios. Others employ simplistic regression models that oversimplify the intricacies of the task while offering limited interpretability. In reality, the pricing of used cars is common but inherently complex, existing research is insufficiently convincing to handle.

This complexity primarily is mainly from the nature of the data, which is typically large, highly unstructured, and inconsistent. To tackle these challenges, we collected data by scraping one of the largest online used car marketplaces, CarMax.com, ensuring the dataset is comprehensive and up-to-date. We systematically extracted both structured and unstructured information, conducted data cleaning process to prepare the data for analysis. Subsequently, we conducted exploratory data analysis (EDA) to uncover patterns and correlations within the data, providing a robust foundation for predictive modeling. We conduct several statistical as well as machine learning approaches to uncover the correlation between variables and target, compared the pros and cons of each approach. These comparative studies not only highlight the performance of different methodologies but also introduce discussions on the trade-offs between

interpretability and predictive accuracy. By evaluating both traditional statistical methods and advanced machine learning algorithms, such as random forests, XGBoost and deep learning methods, this study identifies the most effective approaches for used car price prediction under various conditions.

This project makes main three contributions to the pricing of used cars:

- First, we generate business insights into the factors affecting car resale values through data-driven approaches.
- Second, we provide reliable and transferable statistical models capable of predicting resale values with high accuracy, helping buyers and sellers make rational decisions and promoting transparency, efficiency, and reliability.
- Third, we develop a user-friendly platform incorporating dashboards, predictive tools, and automated reports. Users can easily assess their used car's market value in a way to avoid overpricing or underpricing.

2 Method

2.1 Data Acquisition

The data utilized in this project were primarily sourced from publicly available information scraped from CarMax, one of the most popular online auto trading platforms in North America. To acquire these data, a scraping script was developed using Selectolax for HTML parsing and Playwright as a headless browser. The script was designed to navigate and extract website content, transforming unstructured HTML data into structured formats.

The script includes several key functionalities, such as pagination handling, managing shadow DOM content, error checking for unavailable data, basic data cleaning, and exporting the scraped data into a CSV format. Additionally, the scraping tool allows users to customize the amount of data that is being scraped and define the maximum concurrency, ensuring both flexibility and scalability. Specifically, the script automates the Playwright browser to navigate the search pages of each car brand, extract links to individual car display pages, and scrape detailed information from those pages. The extracted data are stored in a list of dictionaries, which is subsequently converted to a CSV file.

In total, the script successfully scraped 10,367 car information records from 31 different brands, reorganizing the data into a structured format suitable for further analysis.

2.2 Data Cleaning

To refine the dataset and ensure reliability for further development, data cleaning was performed at multiple stages of the project.

Firstly, during the scraping process, the tool included functionality to filter out cars with missing critical information, such as prices or mileage. It also performed data-type and unit conversions during scraping, such as transforming string representations of mileage into numeric formats (e.g., converting "31K" to 31,000).

After acquiring the preliminary dataset through scraping, dedicated cleaning and preprocessing operations were carried out to further refine the data and align it with the project's objectives. Specifically, approximately 12 different cleaning and preprocessing strategies were applied. These included converting textual representations into numerical values for easier handling (e.g., changing "1 owner" to the integer 1), removing records with obvious errors, splitting composite features into separate attributes (e.g., separating "color" into "exterior color" and "interior color"), imputing missing values, and eliminating single-value features to prevent outlier effects. This pre-cleaned dataset served as the foundation for subsequent development.

Furthermore, during the exploratory data analysis (EDA) and model building stages, team members performed further data cleaning tailored to their specific research objectives. For example, features with missing values may be dropped in one analysis but retained in another if they were relevant to the researcher's focus.

2.3 Study of Impact of Mileage

Mileage is widely recognized as a crucial indicator of a used car's resale price. The importance of mileage may also correlate with other features, indirectly influencing the price. These correlations can provide insight into how mileage impacts a car's overall condition, potentially explaining why increased mileage typically leads to decreased prices. Furthermore, understanding these correlations allows the model to estimate a car's mileage range based on other related features when exact mileage data is unavailable. The ability to predict mileage based on provided feature information serves as a key factor in assessing a car's overall condition, especially when mileage data is missing.

2.3.1 Correlation between mileage and resale price

The relationship between a car's mileage and its resale price was examined. Figure 1 illustrates a clear negative correlation between mileage and price across all collected car data, regardless of brand. This outcome is both expected and intuitive.

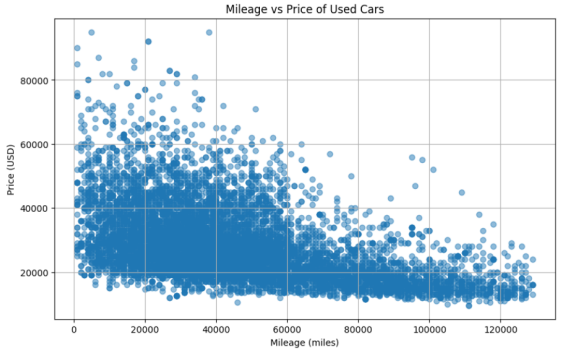


Figure 1: Mileage vs Price of Used Cars

To gain deeper insights, further analysis was conducted to examine the relationship of outliers. A linear regression model was applied to identify car records with abnormal price or mileage values. The relationships between these outliers and other cars are presented in Figure 2. The results indicate that outliers generally exhibit a higher resale value compared to other cars with similar mileage. These outliers will be analyzed further in the next section to uncover potential reasons behind their deviation. From this analysis, the Pearson correlation coefficients for normal cars and outliers were found to be -0.49 and -0.62, respectively, emphasizing the strong influence of mileage on a car's resale value.

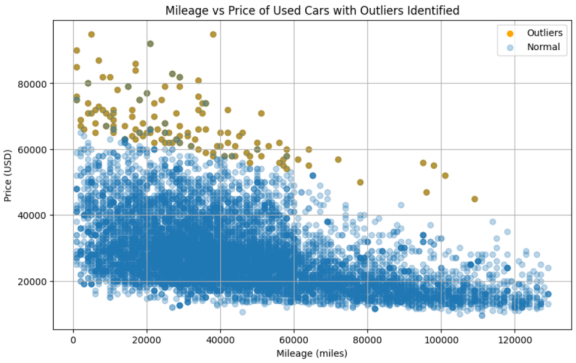
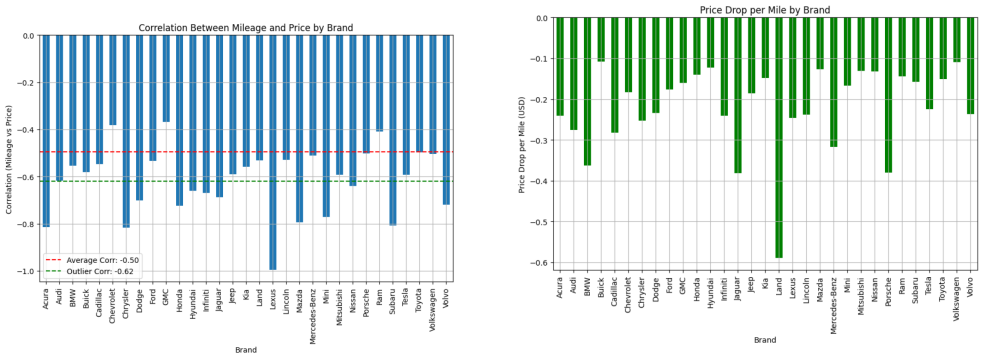


Figure 2: Mileage vs Price of Used Cars with Outliers Identified

2.3.2 Correlation between mileage and Car Brands



(a) Correlation Between Mileage and Price by Brand (b) Price Drop per Mile by Brand

Figure 3: Comparison of Mileage Effects on Price by Brand

The correlations between mileage and resale value for each car brand were further analyzed. The results support our hypothesis that different car brands exhibit varying correlation values. Notably, the price drop per mile differs across brands, highlighting the varying depreciation rates and hedging abilities among them. Additionally, cars recognized as luxury or high-end experience more significant depreciation as mileage increases. Specifically, brands such as Chrysler, Acura, and Lexus show correlation values exceeding -0.8, indicating a strong negative impact of mileage on resale value.

2.3.3 Relationship between mileage and other car features

In this project, more than 20 features were under studying, therefore, To gain further insights about this important feature, it is essential to examine the relationship between mileage and additional features. The correlations found can serve as indicators of how mileage influences a car's overall condition, potentially explaining why increased mileage leads to decreased prices. Additionally, understanding these correlations could enable the model to estimate mileage based on other related features when exact mileage data is unavailable.

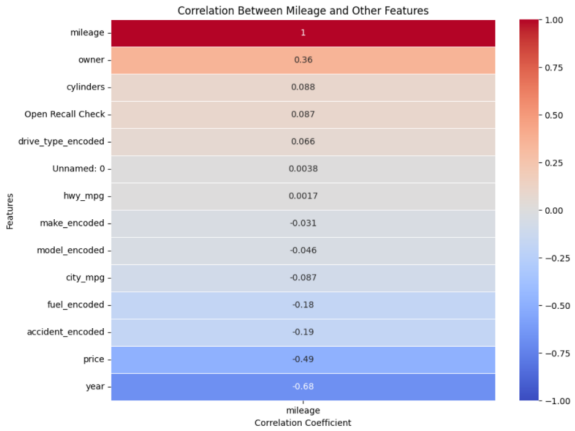


Figure 4: Correlation Between Mileage and Other Features

Specifically, the Pearson correlation coefficients between mileage and other features were calculated, as shown in Figure 4. Features such as VIN and color, deemed less relevant based on intuition, were excluded from the analysis. The results indicate that, apart from price, features with a relatively significant correlation with mileage include year, ownership count, accident history, and fuel type.

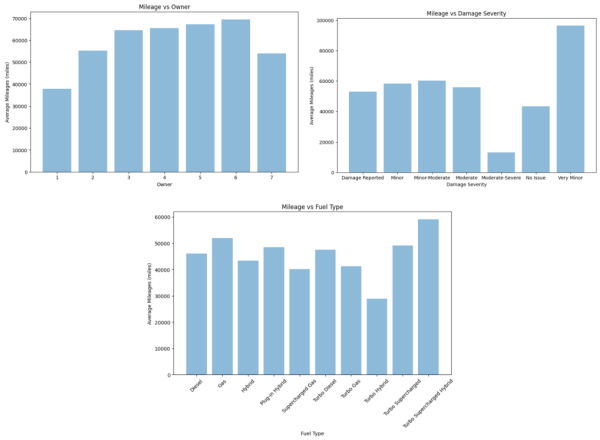


Figure 5: Correlation Between Mileage and Owner, Damage Conditions and Fuel Type

Figure 5 illustrates the relationships between mileage and specific factors, including ownership count, damage severity, and fuel type. The data suggests that mileage generally increases with the number of owners, likely due to more usage across multiple drivers. Regarding damage severity, cars with very minor damage exhibit significantly higher average mileage compared to those with severe damage, indicating that extensive damage might reduce a car's lifespan or usage. For fuel type, average mileage remains relatively consistent among common types like diesel and gasoline, while less common fuel types display greater variation. These findings provides valuable insights into the influence of mileage.

2.3.4 Gradient Boosting Classifier for Mileage Range Prediction

The Gradient Boosting Classifier (GBC) was employed alongside a customized sampling algorithm to predict mileage ranges. The customized sampling algorithm generated additional synthetic data to address underrepresented but potentially significant feature values. This approach was designed to improve the model's understanding of the overall data distribution and mitigate the risk of over fitting.

Gradient Boosting Classifier (GBC) was selected for this task based on the insights obtained from the analysis conducted above. The EDA revealed that certain features, including year, price, make, owner, Damage/Accident, fuel type and owners have varying degrees of correlation with mileage. Among these, although year and price showed the strongest correlations, other features only exhibit moderate to weak correlations (0.1 to 0.4). Given the complex, non-linear relationships among these multiple features,simpler classification algorithms or models such as decision trees or logistic regression might struggle to capture the underlying pattern effectively. GBC, an ensemble method that combines multiple weak learners to form a robust model, was therefore deemed as a good choice for this task [4].

The original dataset exhibited an underrepresented bias in certain important feature values. Specifically, some critical values lacked sufficient data, which could hinder the model's performance on unseen real-world data and increase the risk of over fitting. To address this issue, a customized algorithm named generate_uncommon was developed to generate synthetic data for underrepresented feature values. Instead of treating these limited feature values as outliers and discarding them, the algorithm increased their representation in the dataset. This ensured that the model was exposed to a more balanced distribution of data, particularly for features that might be crucial for predictive accuracy.

The algorithm divided features into two categories, categorical and numerical, and applied tailored methods for each. For categorical variables, such as car make, the percentage distribution of each value in the original dataset was calculated, and synthetic rows were generated to maintain these proportions. For instance, if 10% of the cars in the original dataset were Honda, then 10% of the synthetic rows also had their car make set to Honda. For numerical variables, such as mileage and price, the algorithm assumed a normal distribution for each feature, calculating their mean and standard deviation. Synthetic values were then sampled from these distributions to approximate real-world patterns.

The model's performance was optimized by fine-tuning key hyper parameters such as n_estimators, learning_rate, max_depth, and min_samples_split. The GridSearchCV tool was employed to systematically test various combinations of these parameters and identify the optimal configuration. This approach facilitated the creation of the best possible model within the defined parameter search space in an efficient and methodical manner.[5]

2.4 Study of Impact of Brand

2.4.1 Correlation between Brand and Resale Price

We examined the impact of brand recognition on the resale price of used cars. Specifically, we hypothesized that certain brands generally have higher average resale prices, with luxury and popular brands commanding a premium in the market. To test this hypothesis, we grouped the data by brand, and the average resale price for each brand was calculated.

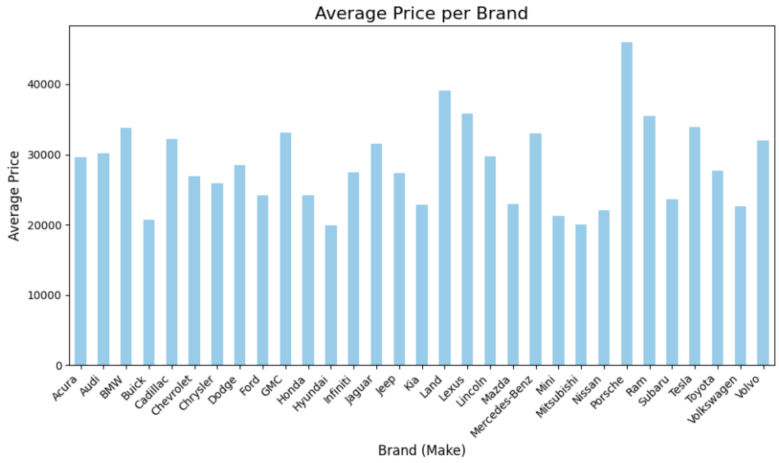


Figure 6: Average Price of Used Cars Grouped by Brand

As shown in Figure 6, the analysis revealed a significant variation in average resale prices among different brands. The top five brands with the highest average resale prices were Porsche, Land Rover, Lexus, Ram, and Tesla, which are well-known luxury or popular brands. On the other hand, the bottom five brands—Hyundai, Mitsubishi, Buick, Mini, and Nissan—had the lowest average resale prices, this indicating that their position as more affordable or economy-focused brands.

2.4.2 Deprecation Rate of Different Brands

To further analyze how different brands depreciate, we conducted a cohort analysis. While brands significantly influence resale prices, they show varying price variances and depreciation rates over time. The top 10 brands with the highest variance include luxury brands like Mercedes-Benz, Porsche, and BMW, as well as mass-market brands such as Chevrolet and Ford. Luxury brands typically exhibit large price differences due to their diverse offerings, from entry-level to high-end models, while mass-market brands produce a wide range of vehicles, leading to broader price variance.

Figure 7: Price Distribution of Top 10 Brands with Highest Variance

Figure 7 highlights that Mercedes-Benz and Chevrolet have the most outliers, with luxury brands like Land Rover and Porsche showing outliers due to high-end or limited-edition models. Similarly, mass-market brands like Ford and Chevrolet show outliers due to their diverse lineups, including sedans, SUVs, and trucks. These outliers, though extreme, remain essential to resale price trends and are not excluded from analysis.

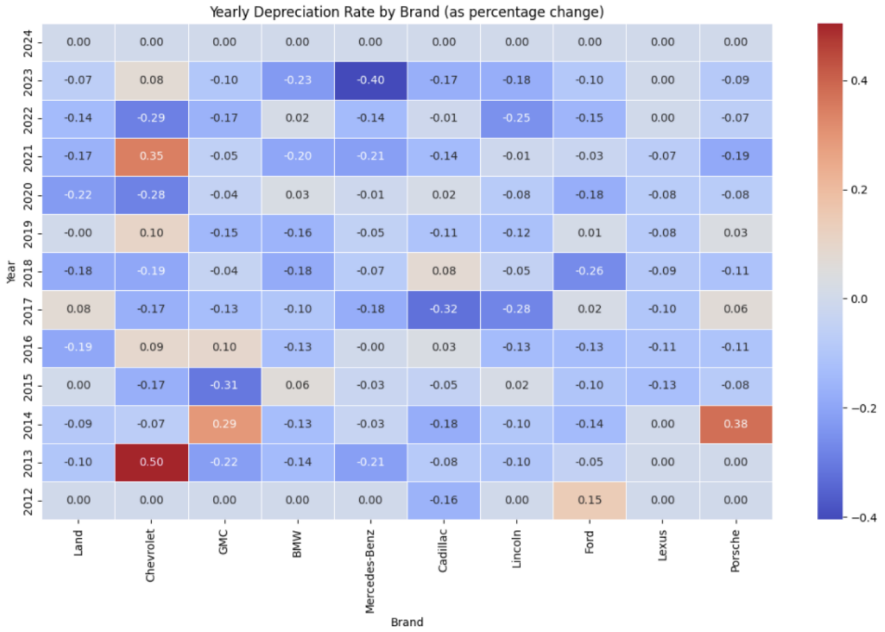


Figure 8: Yearly Depreciation Rate by Brand(as percentage change)

A heatmap of depreciation rates from 2012 to 2021 shows most brands experienced reduced depreciation or price recovery after 2020, especially Land Rover and GMC. Unusual cases, such as Chevrolet and Ford, showed price increases during specific periods, likely due to supply-demand fluctuations or marketing strategies. These findings underscore the complexity of brand-specific depreciation trends and market dynamics.

2.5 Study of Impact of Owners

2.5.1 Correlation between Owners and Resale Price

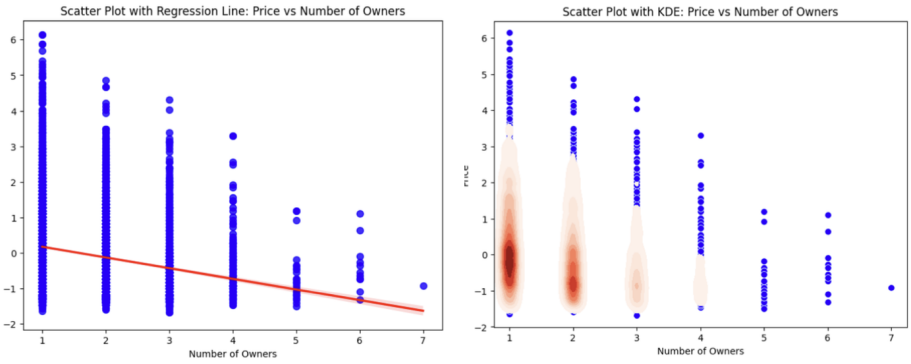


Figure 9: KDE and Regression

As shown in Figure 9, there is a significant negative correlation between the number of owners and resale price. Vehicles with fewer owners (e.g., 1 or 2) tend to have higher prices, with prices more concentrated in the higher range. In contrast, vehicles with more owners show lower and more dispersed price distributions. However, this trend is not absolute, as vehicles with fewer owners can also exhibit greater price variance due to other influencing factors.

2.5.2 Price Sensitivity to Owners Across Brands

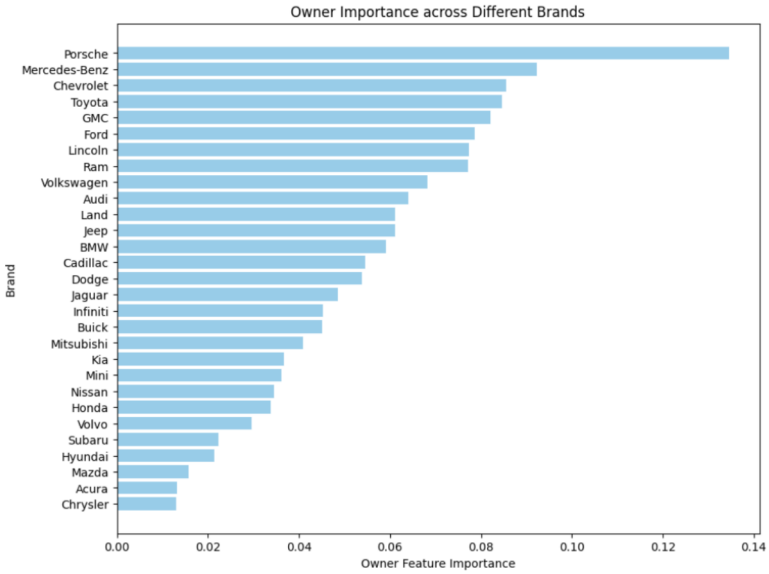


Figure 10: Owner Importance across Brands

Different brands exhibit varying price sensitivity to the number of owners. Luxury brands like Porsche and Mercedes-Benz show significantly higher sensitivity, as buyers of these vehicles often prioritize ownership history. In contrast, non-luxury brands such as Chrysler and Mazda display lower sensitivity, with buyers focusing more on factors like mileage and condition. Middle-class brands such as Toyota and Chevrolet exhibit moderate sensitivity, maintaining good resale value and balanced price retention.

However, training models exclude brands with fewer than 100 samples, which can lead to bias. For example, Lexus, a luxury brand, is excluded from training but shows high sensitivity to the number of owners when included, generating the highest feature importance for this variable. This highlights the need for more balanced samples to improve the accuracy of feature importance across different brands.

2.6 Study of Impact of color

2.6.1 Americans' Preferences for Car Colors

Color is a characteristic that is difficult to intuitively say how important it is to the price of a car, because it depends largely on people's preferences and the design of the car itself. Such a confusing factor is worth studying its impact on the price of a car.

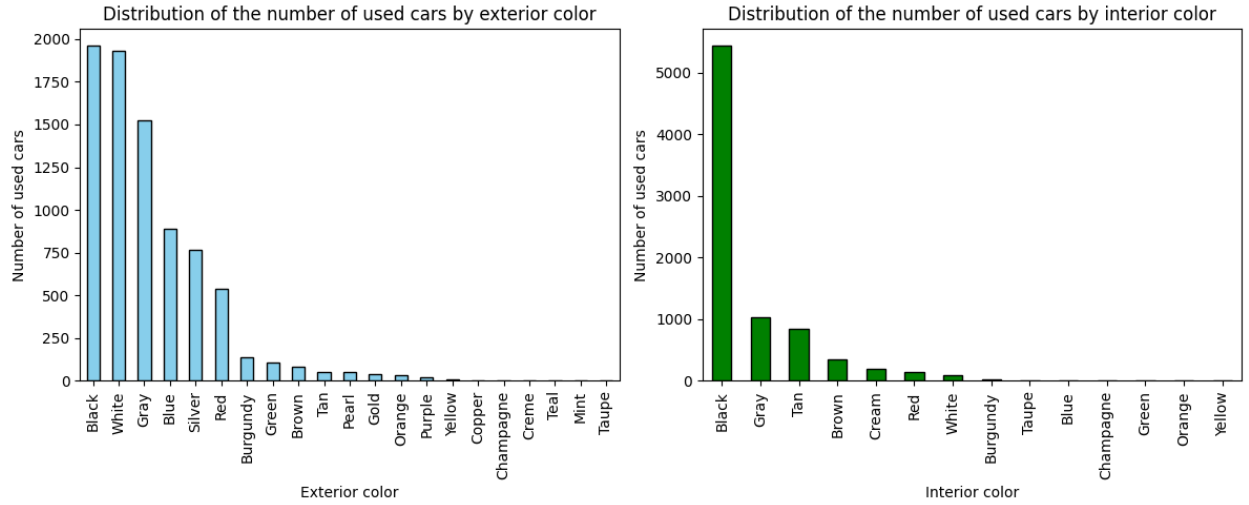


Figure 11: Distribution of the number by interior color and exterior color

As is shown in Figure 11 Americans exhibit a clear preference for neutral, timeless car colors, with white, black, and silver leading exterior choices and black dominating interior options. These trends emphasize practicality, broad appeal, and market demand for standard color palettes.

2.6.2 Relationship between Color Preference and Manufacturer and Year

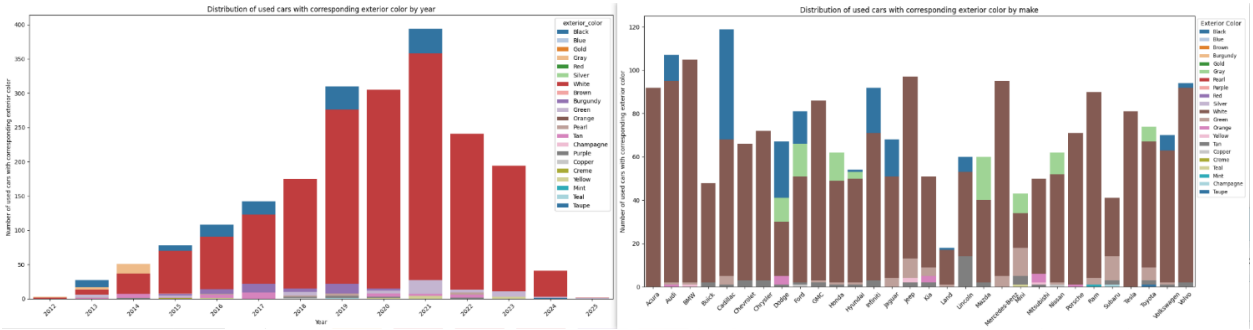


Figure 12: Distribution of used cars with corresponding exterior color by make and year

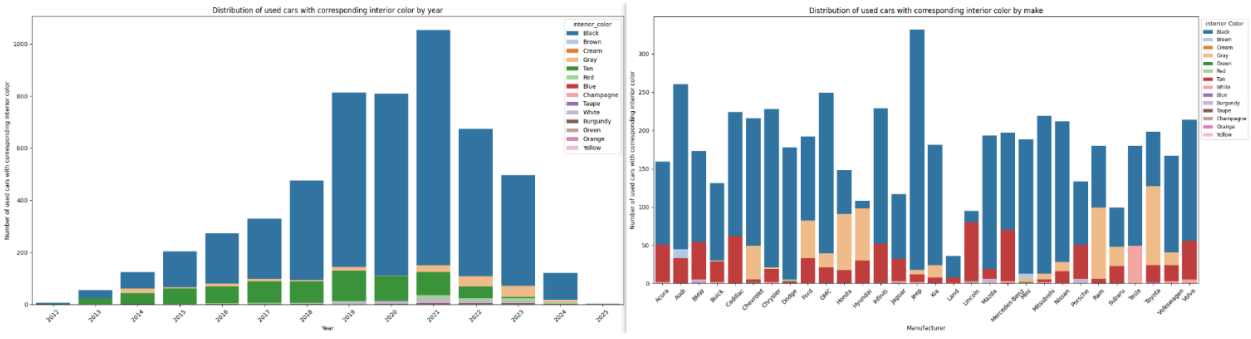


Figure 13: Distribution of used cars with corresponding interior color by make and year

As shown in Figure 12, the analysis of the dataset reveals that while the exterior color of used cars, particularly white, is predominantly consistent across different years and manufacturers, as illustrated in the accompanying graphs, there is

no discernible relationship between exterior color and car price, suggesting that exterior color has minimal impact on pricing due to its concentrated distribution and the general lack of price differentiation among popular colors.
In Figure 13, we know that the interior colors of used cars are concentrated on black, gray, and tan. As shown in the graph above, as the years progress, most used cars have black interiors. This trend is evident at the right in Figure 13, which shows the distribution of interior colors across different manufacturers, although Honda and Hyundai have some different preferences.

2.6.3 Price Sensitivity to Different Colors

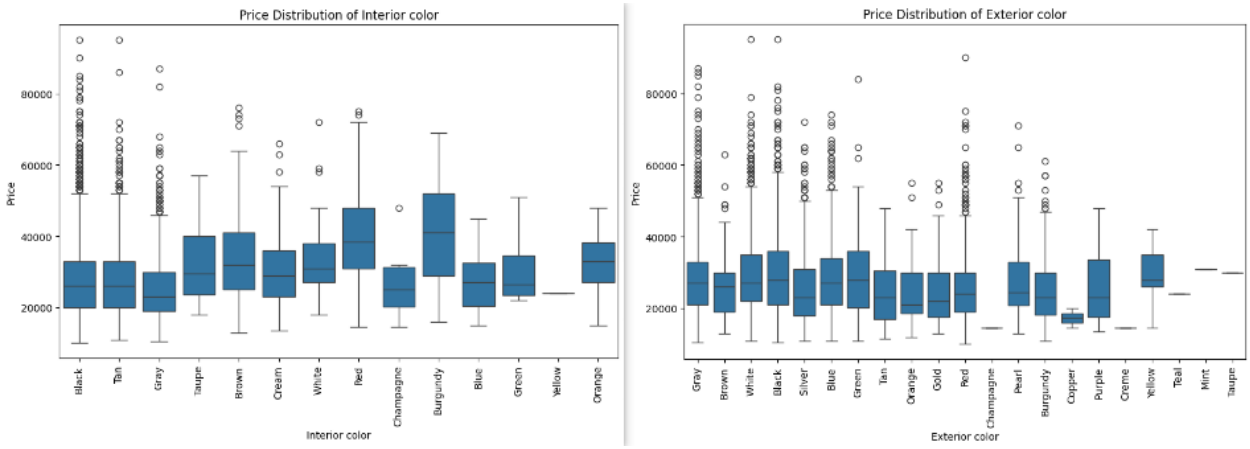


Figure 14: Price Distribution of color

Based on the Figure 14, the exterior color of used cars shows minimal influence on their prices, as the distribution is heavily concentrated, and popular colors such as white, black, and silver do not exhibit a significant association with lower prices. In contrast, the interior color appears to have a modest impact, with red and burgundy interiors correlating with relatively higher car prices.

2.6.4 Feature Selection and Model Analysis for Price Prediction

Feature Selection Using Statistical and Machine Learning Methods The SelectKBest algorithm[18] was employed to identify the most influential features for predicting the price of used cars. SelectKBest utilizes statistical tests, such as the Chi-Square test[15], ANOVA F-test, or mutual information score, to rank features based on their statistical association with the target variable. The inclusion of color as a selected feature indicates that it may play a significant role in price prediction; however, its exclusion suggests that color (whether exterior or interior) is not a primary determinant.

In addition to statistical methods, Random Forest Classification[18] was applied to evaluate the predictive power of various feature subsets, including color alone, all features, and subsets excluding color. Leveraging the ensemble nature of Random Forest, the number of decision trees (n_estimators) was tuned to optimize model accuracy. Random Forest also provides feature importance scores, offering insights into the key predictors influencing price.

SelectKBest identifies the top k features based on statistical scores, whereas Random Forest ranks features based on their importance as determined during model training. Notably, the feature rankings produced by these two methods diverge. To validate and identify the most meaningful features, Linear Regression was employed as an additional measure, with the models evaluated based on R^2 and Root Mean Square Error (RMSE) metrics. This triangulation approach provides a robust evaluation of the algorithms' effectiveness in selecting critical features.

The primary tunable parameter in the Random Forest Classifier is the number of decision trees (n_estimators). Increasing this parameter typically enhances model performance until a saturation point is reached. The random_state parameter was fixed across experiments to ensure reproducibility and allow for a consistent comparison of how different feature sets (e.g., color only, all features, all except color) influence model accuracy. Optimal values of n_estimators were identified using GridSearchCV[5], which systematically evaluates the parameter's impact on model performance.

Given the discrepancy between the feature rankings from SelectKBest and Random Forest, Linear Regression was utilized to assess which method identified the most impactful subset of features for price prediction.

2.7 Study of Impact of Fuel

2.7.1 Correlation between Fuel Type and Other Features of used cars

Fuel type is a key factor in determining a car's resale value, our analysis focused on how fuel type impacts resale value across different car brands. To deepen our insights, it's crucial to examine the relationship between fuel type and other vehicle attributes. By understanding these correlations, we can enhance the model's ability to estimate fuel type based on related features, even when direct fuel type data is missing. This study focuses on a categorical analysis by examining two non-numeric and two numeric parameters.

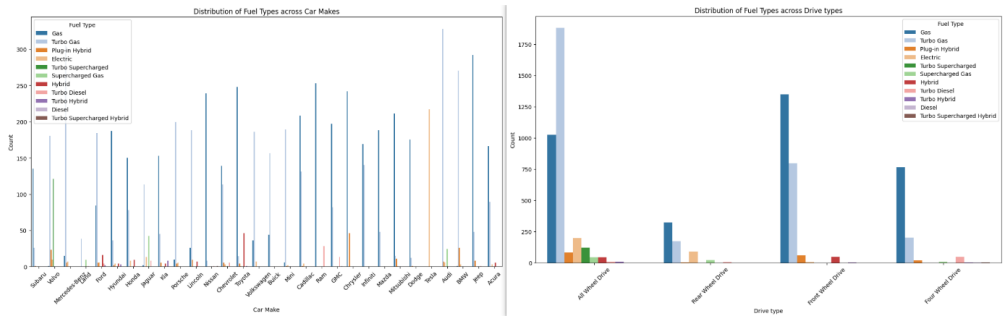


Figure 15: Distribution of Used Cars with Fuel by Make and Drive Type

The Figure 15 reveals that Gas and Turbo Gas are the dominant fuel types across various brands and drive types. However, the distribution of all four fuel types varies considerably by drive type, with certain brands, such as Volvo, demonstrating a preference for specific fuel types.

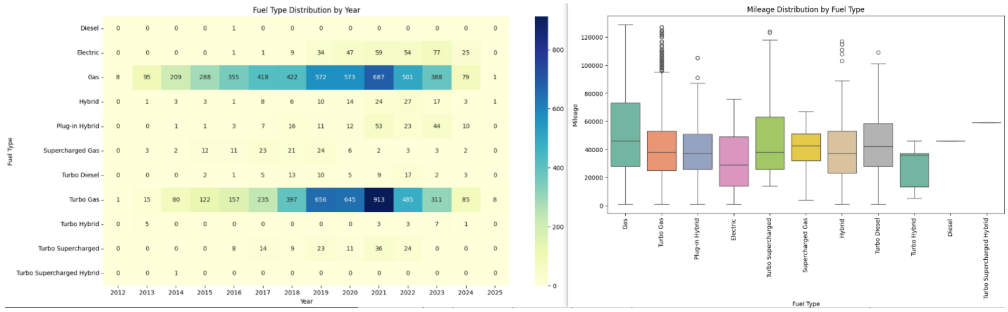


Figure 16: Distribution of Used Cars with Fuel by Year and Mileage

The findings in Figure 16 also indicate that Gas and Turbo Gas have remained consistently prevalent fuel types in the used car market over time, suggesting a weak correlation between vehicle production year and fuel type. Additionally, the median mileage is comparable across all fuel types, implying that mileage is not significantly influenced by fuel type.

	Feature	p-value
0	year	9.901719e-102
1	make	0.000000e+00
2	model	0.000000e+00
3	price	1.523378e-128
4	mileage	2.614113e-73
5	Miles per gallon	0.000000e+00
6	Transmission	4.490467e-01
7	owner	1.375909e-14
8	VIN	4.891056e-01
9	class	0.000000e+00
10	Auction Brand / Issues	9.999797e-01
11	Accident / Damage	2.908930e-04
12	Open Recall Check	0.000000e+00
13	Odometer Check	7.506976e-01
14	Certified Pre-Owned	8.499720e-04
15	cylinders	NaN
16	Drive type	1.900100e-305
17	Miles per gallon equivalent (MPGe)	1.150855e-45
18	Range (when new)	3.358964e-55
19	Time to fully charge battery (240V)	2.049358e-66
20	Motor	0.000000e+00
21	Bed Length	9.047790e-22
22	exterior_color	1.645491e-02
23	interior_color	2.483810e-171

Figure 17: Model Analysis of Fuel and Other Features

To further investigate the relationship between fuel type and other features, a Chi-Square test[15][16] was applied to categorical features as is shown in Figure 17, while ANOVA was utilized for numeric features. Features with a p-value below 0.05 were deemed significantly associated with fuel type. The resulting significant features include make, model, miles per gallon, class, open recall check, drive type, bed length, and interior color.

2.7.2 Hyperparameter Tuning and Feature Importance Analysis for Fuel Type Prediction

GridSearchCV[5] was employed to fine-tune the hyperparameters of the CatBoostClassifier, optimizing parameters such as iterations, learning_rate, depth, random_seed, and verbose. This process ensured optimal model performance for predicting fuel type. The CatBoostClassifier[17] was further utilized to analyze relationships between fuel type and other features, providing both predictions and feature importance rankings. While using CatBoostClassifier, we also found that CatBoostRegression[17] is also a good way to predict prices of used cars.

GridSearchCV systematically optimized key hyperparameters of the CatBoostClassifier, including iterations, depth, and learning_rate, to maximize accuracy. The CatBoostClassifier not only predicted fuel type but also ranked features by their importance, offering insights into the relationship between fuel type and other variables. This approach ensured both predictive accuracy and interpretability of results.

2.8 Study of Impact of Accident History

2.8.1 Price pattern is consistent

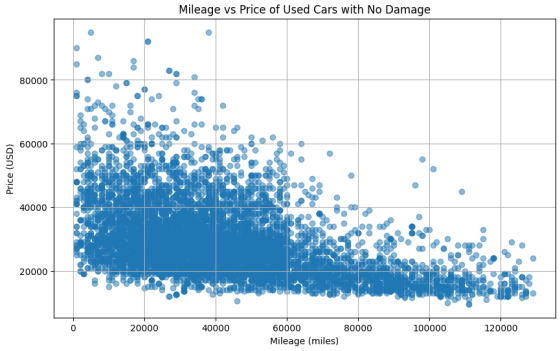
The used cars with different damage levels share a similar pattern between mileage and price: We plot the figure between the mileage and price of different accident history groups: no damage, minor damage, moderate damage, and general damage in Figure 18. We can observe that used cars with different damage levels maintain a relationship between mileage and price. We further check it by computing the Pearson correlation scores as following:

- Pearson correlation of used car with no damage: -0.47140781824890604.
- Pearson correlation of used car with minor damage: -0.5117868757602223.
- Pearson correlation of used car with moderate damage: -0.5172322127037858.
- Pearson correlation of used car with damage reported: -0.520546544670029.

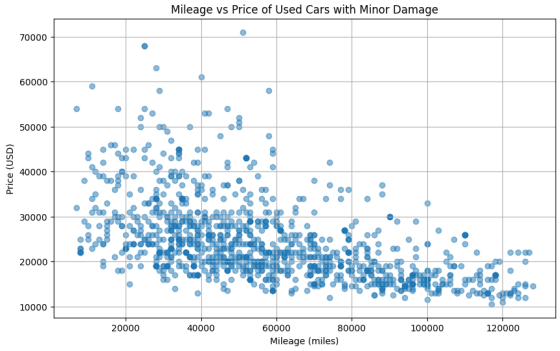
The above Pearson correlation results verify that used cars with different damage levels share a similar pattern between mileage and price.

2.8.2 Accident history has negative impacts on resale price

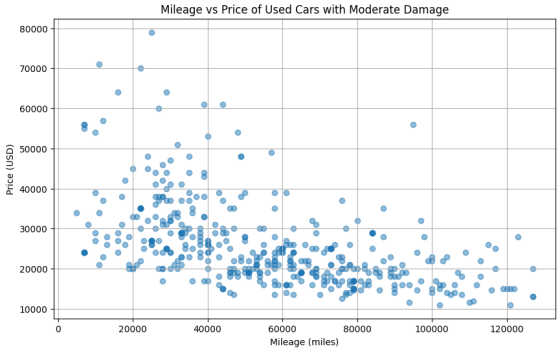
Used cars with higher damage level tend to have lower price: We conduct the linear regression between mileage and resale price on the above 4 groups: no damage, minor damage, moderate damage, and damage reported. From Figure 19, we can observe that in most mileage ranges, used cars with damage will have lower resale prices than cars with no damage/accident and higher damage levels will have more negative impact on the resale prices.



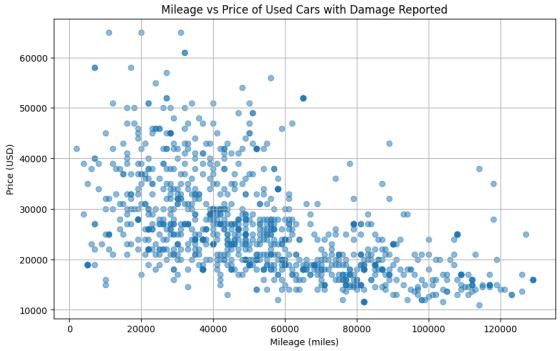
(a) No damage



(b) Minor damage



(c) Moderate damage



(d) General damage

Figure 18: Mileage and price in different accident history groups

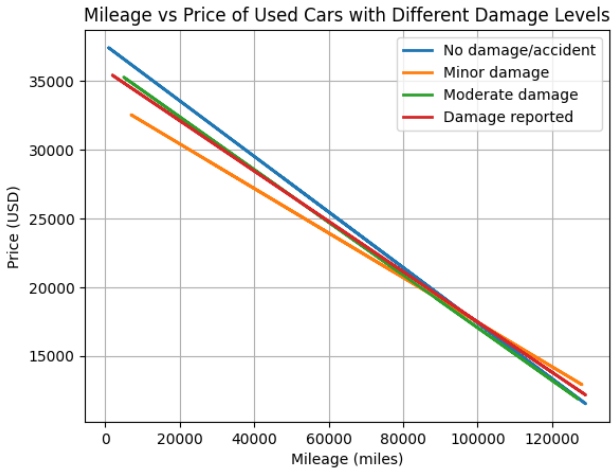


Figure 19: Linear regression results between mileage and resale price on 4 groups

2.9 Study of Impact Variance of Accident History between Different Makes

2.9.1 Price pattern is consistent across different car makes

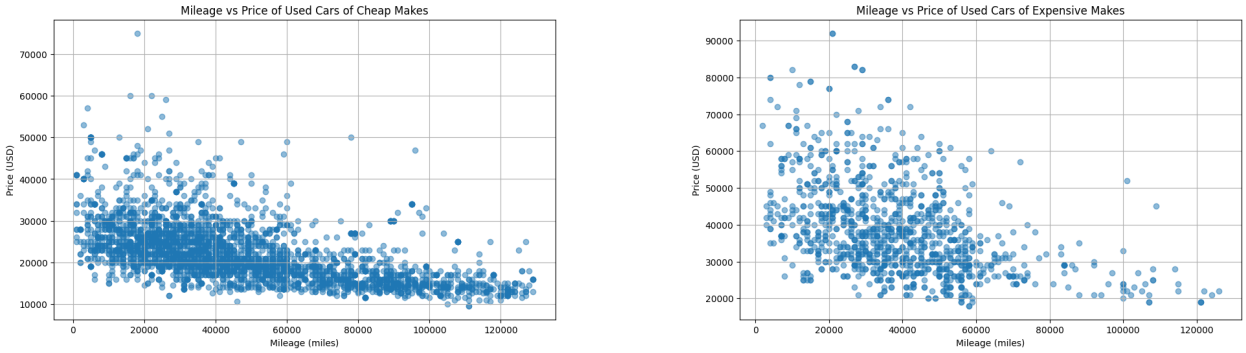
We first construct two make groups: the cheap car group and the expensive car group. Their statistics are shown in Table 1 and Table 2. We plot the figure between the mileage and price of these two different make groups in Figure 20.We can observe that used cars with different makes maintain a similar relationship between mileage and price. We further check it by computing the Pearson correlation scores as following:

Table 1: Cheap car group statistics

Make	Average resale prices ($)
Hyundai	19954.47
Mitsubishi	20048.48
Buick	20720.97
Mini	21246.32
Nissan	22001.95
Volkswagen	22651.10
Mazda	22927.38
Kia	22931.19
Subaru	23680.25
Ford	24181.82

Table 2: Expensive car group statistics

Make	Average resale prices ($)
Porsche	45998.00
Land Rover	39102.95
Lexus	35798.00
Ram	35472.19
Tesla	33868.68



(a) Cheap makes

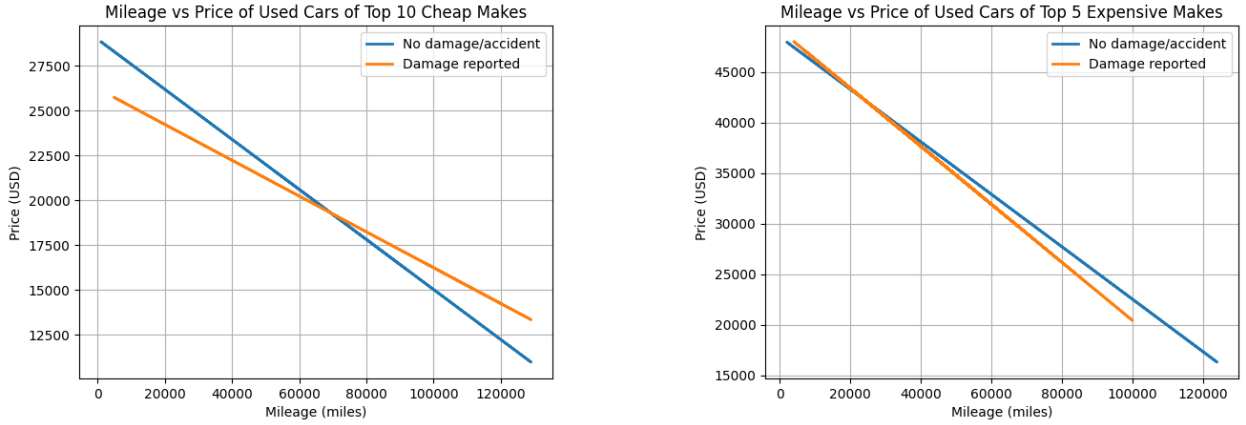
(b) Expensive makes

Figure 20: Mileage and price in different make groups

- Pearson correlation of used car of expensive makes: -0.4487958243732463.
- Pearson correlation of used car of cheap makes: -0.5643709389958534.

The above Pearson correlation results further verify that the used cars with different makes share a similar pattern between mileage and price.

2.9.2 Makes' sensitivities on resale price towards the accident record



(a) Cheap makes

(b) Expensive makes

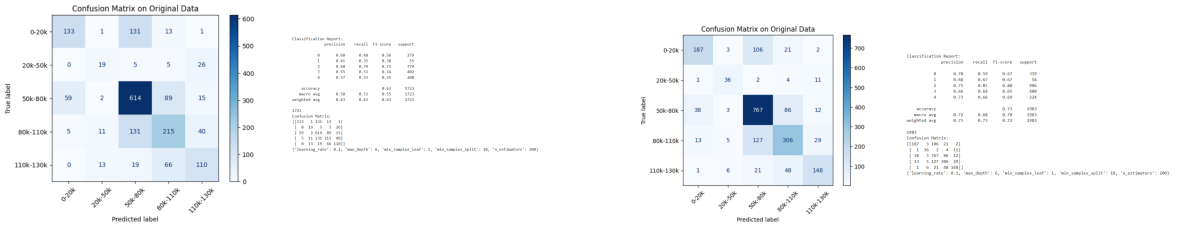
Figure 21: Linear regression results for the two make groups with different accident records

Used cars with different brands will have different sensitivity on price towards the accident/damage record: We conduct linear regression on the above two groups of data shown in Figure 21. Since most used cars on sale have mileage smaller than 80000 (1114/107 for samples in the plot). Thus we focus on this range. We can observe that used cars with cheap auto makes are more sensitive to whether the used cars have been through accidents (the used car with an accident reported has a much lower price.) But for used cars of expensive makes, the effect from accidents is much more slight. It indicates that cheap make's car prices are more sensitive towards accident records.

3 Results

3.1 Gradient Boosting Classifier for Mileage Range Prediction

The Gradient Boosting Classifier (GBC) was employed alongside a customized sampling algorithm to predict mileage ranges. The customized sampling algorithm generated additional synthetic data to address underrepresented but potentially significant feature values. This approach was designed to improve the model's understanding of the overall data distribution and mitigate the risk of over fitting. The model achieved an accuracy of 63% on the original dataset, which increased to 73% when synthetic data was incorporated.



(a) Mileage Range Predictor on Actual Data

(b) Performance of Mileage Range Classifier on Synthetic Data

Figure 22: Comparison of Mileage Range Prediction on Actual and Synthetic Data

The feature importance graph reaffirmed the findings from the EDA stage, with the ranking of feature importance closely aligning with their correlation values. A notable observation is that while the EDA showed a stronger correlation between year and mileage compared to price, the Gradient Boosting Classifier (GBC) identified price as the most influential feature. This discrepancy suggests that year, as a standalone feature, might not be as representative, serving instead as an indirect measure of a car's condition. Furthermore, the strong relationship between mileage and price reinforces mileage's importance in predicting resale value.

From the confusion matrix and classification report, the model's accuracy of 63% on the original dataset reflects moderate performance in capturing the complex relationships within the data. However, the macro-averaged recall and F1 score indicate variability in the model's performance across different mileage categories. Specifically, the low precision for class 0 (0k-20k) and class 1 (20k-50k) suggests the model struggled to correctly identify the characteristics of these categories. This is evident in the confusion matrix, where class 1 and class 2 (50k-80k) were frequently misclassified. Conversely, class 2 achieved the best precision and recall, likely due to having a larger number of data points in this range, which may also explain why the model performed better with the synthesized dataset.

When synthetic data was incorporated, the model's performance improved across all five mileage categories. This improvement highlights the utility of additional data in enhancing model performance and suggests that the sampling generation algorithm effectively reinforced important patterns without introducing significant confusion. However, as the synthetic data was generated based on the distribution of the original data, the observed improvement in accuracy might partly result from overfitting or capturing more granular details within the existing patterns. Further investigation is required to validate the algorithm's effectiveness and ensure its generalizability.

3.2 XGBoost with LSTM neural network for Brand Analysis

The experiment utilized XGBoost[8] and LSTM[10] models to analyze whether a brand's resale price is influenced by the prices of competing brands. First, the average resale price for each brand by year was calculated to construct a structured dataset. An XGBoost model was trained to predict the resale price of a target brand using the average resale prices of all other brands as input features. From the trained XGBoost model, feature importance scores were extracted to identify the top three brands with the greatest influence on the target brand's resale price.

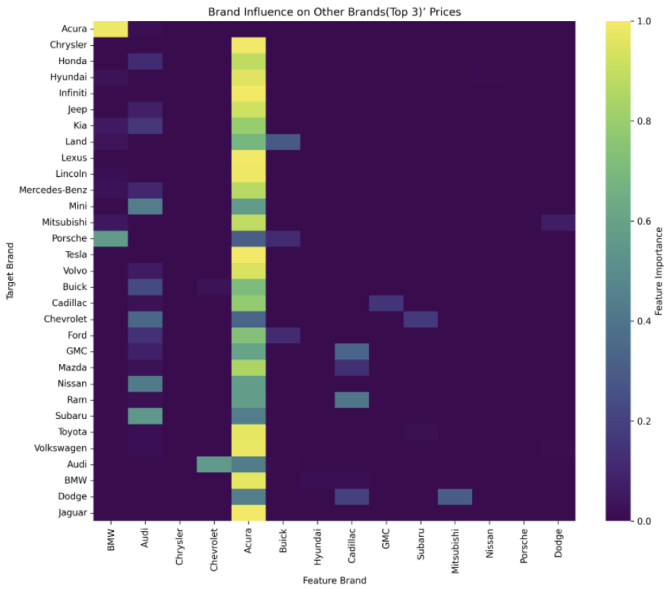


Figure 23: Top 3 Most Significant Brands to Target Brand

Figure 23 confirmed logical relationships in most cases. For example, Porsche was most influenced by BMW, while Mercedes-Benz showed strong correlations with Audi and Acura. Acura, however, appeared to have a disproportionate impact on many other brands, likely due to its overrepresentation in the dataset. This underscores the need for balanced sampling when analyzing feature importance.

XGBoost[9], a gradient boosting algorithm well-suited for structured data, was employed to extract the feature importance of competing brands. The XGBRegressor was trained on the average price data of all brands, targeting the resale price of each brand. The top three brands with the highest feature importance were identified as the most relevant competitors influencing the target brand's resale price. For instance, the analysis revealed that Porsche is most influenced by BMW, while Mercedes-Benz is strongly influenced by Acura and Audi. Interestingly, Acura emerged as a significant influencer for many brands, likely due to its substantial representation in the dataset compared to other brands. This observation emphasizes the importance of balanced datasets in deriving unbiased insights.

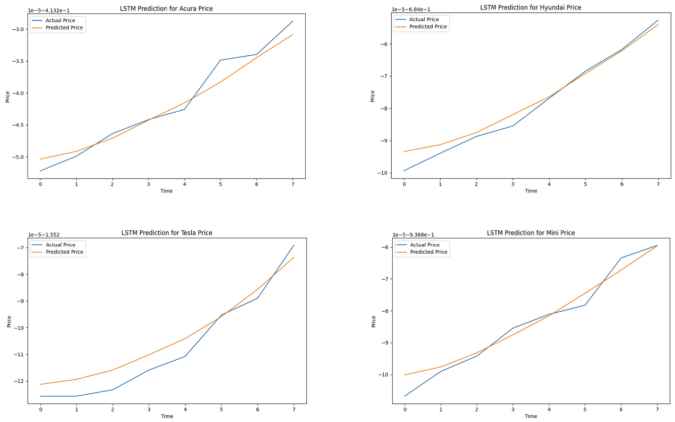


Figure 24: Sample Brands LSTM Prediction Over Years(log scale with x-axis as time step)

While XGBoost identified influential brands, it only provided a static view of brand influence. To account for temporal dependencies and lag effects, an LSTM[11] neural network was employed for time series prediction. The LSTM

model was trained on sequential data constructed using a sliding window approach, where the prices of the top three influencing brands from previous years were used as input features to predict the target brand's price in the current year. The MinMaxScaler was applied to normalize the data for efficient training, and the model's loss history was recorded to monitor its learning progress over epochs.

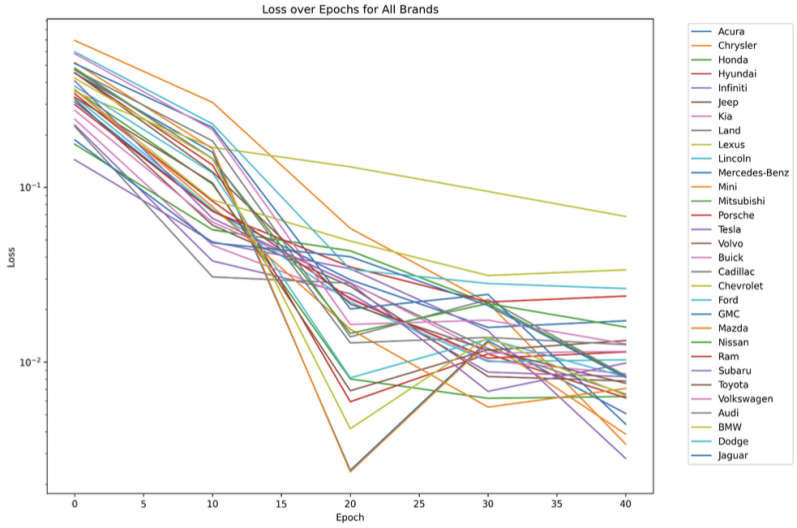


Figure 25: Loss History for Each Brand

LSTM effectively captured temporal trends in resale prices, demonstrating its suitability for analyzing sequential data. The model achieved a significant reduction in training loss after 50 epochs, and its predictive performance was evaluated using metrics such as R^2, Mean Absolute Error (MAE), and Root Mean Squared Error (RMSE). The R^2 score of approximately 0.97 indicates the model's high accuracy. After rescaling the predictions to the original price range, the MAE and RMSE values, confirmed that the combination of XGBoost for feature selection and LSTM for time series prediction outperformed single-model approaches.

The LSTM model successfully captured year-over-year trends in resale prices across brands, demonstrating its robustness in predicting future values. A sample plot (log scale) of LSTM predictions over time highlights the model's ability to track price fluctuations and provide reliable forecasts. Compared to using a single model, the combined approach of XGBoost for feature selection and LSTM for time series forecasting significantly reduced prediction errors, as evidenced by lower MAE and RMSE values.

3.3 SelectKBest and Random Forest Model for Feature Selection

Figure 26 shows the result of top 10 important encoded features selected by SelectKBest model[18], which can be ranked by features' scores. Specifically, when analyzing the top 20 features, neither exterior nor interior color emerged as a significant predictor of price.

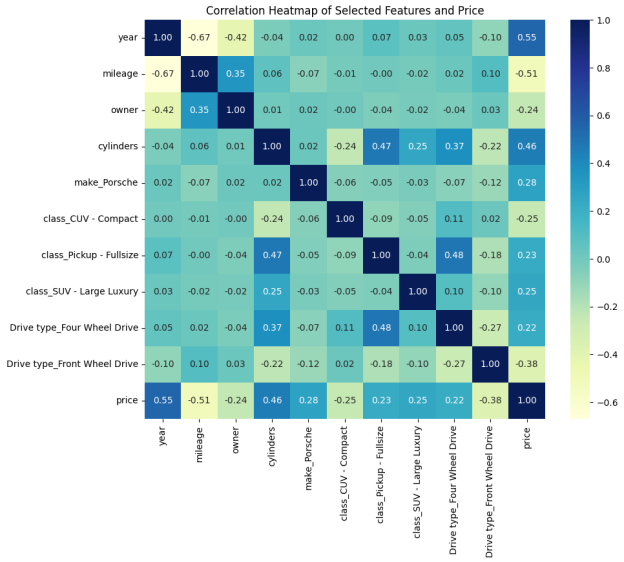


Figure 26: Correalationship Heatmap of Important Features and Price.png

Linear Regression experiments, as is shown in Figure 27, reveal that Random Forest outperforms SelectKBest in identifying the top k features.The Random Forest model demonstrated strong performance, achieving a low RMSE of 3427.9 and an R^2 value of 0.893. This indicates that Random Forest[18] is a robust choice for price prediction tasks, particularly when paired with an optimized feature set identified through its inherent feature importance measures.

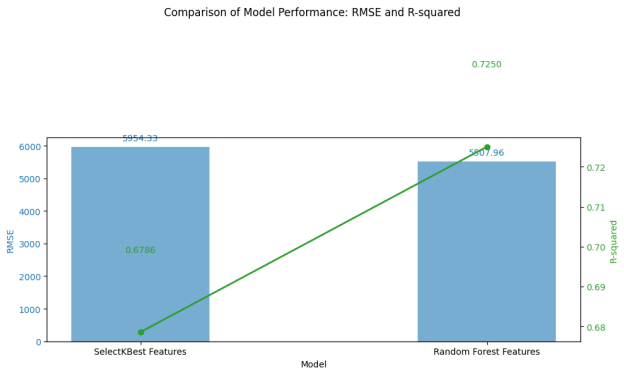


Figure 27: Comparison of Model Performance-RMSE and R-squared

As is shown in Figure 28, Models trained with all features, as well as those excluding color, performed comparably, with a slight performance advantage observed when color features were excluded. These findings suggest that color does not substantially contribute to price prediction and may be omitted without compromising model accuracy.

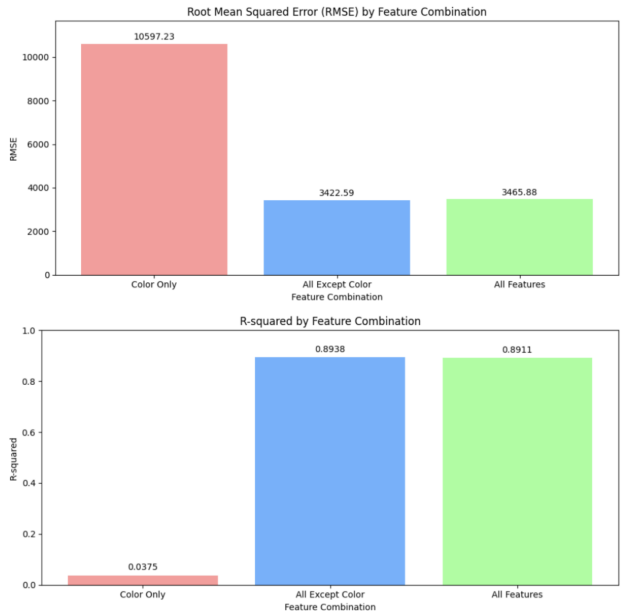


Figure 28: RMSE and R-squared by Feature Combination

The analysis confirms that Random Forest is the most suitable model for predicting used car prices, offering high accuracy and reliable feature importance measures. Furthermore, the negligible impact of color features on model performance supports their exclusion from the feature set. This study underscores the value of combining statistical methods, machine learning models, and linear regression for feature selection and model evaluation in predictive analytics.

3.4 CatBoostClassifier and CatBoostRegressor for Non-Numeric and Numeric Predictions

Based on exploratory data analysis (EDA), eight features with low p-values—make, model, Miles per gallon, class, Open Recall Check, Drive type, Bed Length, and interior_color—were selected as input variables. Using these features, the CatBoostClassifier[18] achieved an accuracy of 0.913 in predicting fuel type. This high accuracy underscores the effectiveness of the model in utilizing highly relevant features for classification. During the analysis of the CatBoost model, it was unexpectedly observed that the CatBoost regression model[18] demonstrates high accuracy in predicting used car prices. Specifically, when the top k most important features are used as inputs, the model achieves a prediction accuracy exceeding 90% within a price deviation range of 10%. In comparison, the performance of the Random Forest model is slightly less consistent under similar conditions.

The CatBoostClassifier's feature importance analysis revealed that Drive type, class, and make were the most influential predictors of fuel type. Training the model exclusively with these three features resulted in an accuracy of 0.826 in Figure 29, indicating their strong predictive power. This reduction in input dimensions demonstrates that these features alone are sufficient to make reasonably accurate predictions.

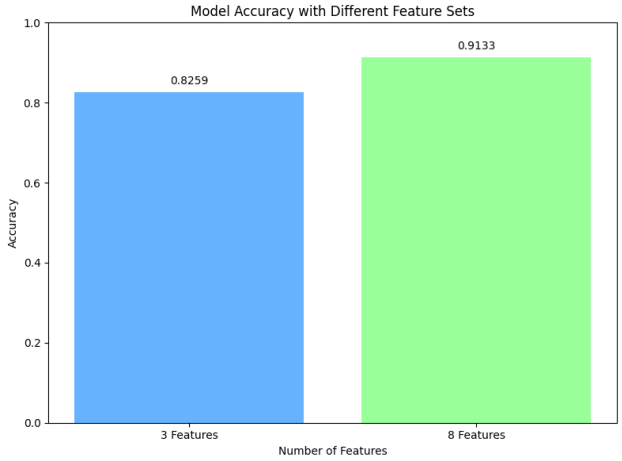


Figure 29: Model Accuracy with Different Feature Sets

An analysis of error distribution in Figure 30 highlighted that the model performs well for larger fuel type categories, such as Diesel and Turbo Diesel, where data instances are more frequent. However, smaller classes, such as Gas and Supercharged Gas, showed weaker performance, suggesting the need for additional data or more refined feature engineering to enhance differentiation between these categories. Additionally, the model exhibited some confusion between similar fuel types, such as Diesel and Turbo Diesel, indicating that further exploration of additional features or advanced feature engineering could improve class separation.

Figure 30: Enter Caption

3.5 Neural Network for Binary Accident prediction

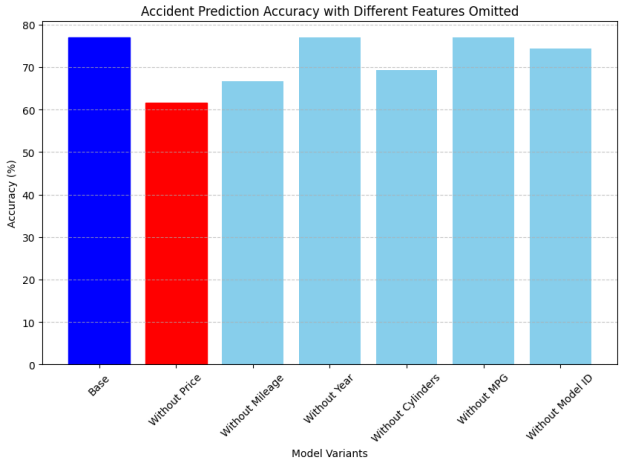


Figure 31: Accident Prediction Accuracy with Different Features Omitted

In this part, we show that the used cars' price and accidents have the strongest correlation by conducting classification to predict whether the car has at least 1 accident. We build a neural network with hidden dimension 100 and conduct experiments on Mercedes Benz's car price data.

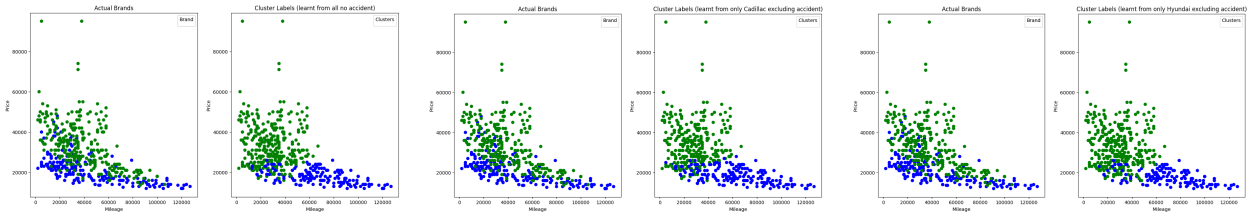
Data processing : We convert 'Miles per gallon' string data into the average data between '$city$' and 'hwy' and convert '$model$' string data into the '$model_{id}$' number value to model this attribute as different models of a car brand have different beginning prices.

Model variant : We build neural network classifiers with hidden dimension 100 with following different inputs:

1. Base experiment with all attributes '$mileage$', '$year$', '$cylinders$', '$price$', '$AverageMPG$' and '$model_id$' as the input attribute and make the prediction on accident classification (no accident: 0, at least 1 accident: 1)
2. Ablation experiment without '$price$' attribute fed in.
3. Ablation experiment without '$mileage$' attribute fed in.
4. Ablation experiment without '$year$' attribute fed in.
5. Ablation experiment without '$cylinders$' attribute fed in.
6. Ablation experiment without '$AverageMPG$' attribute feeded in.
7. Ablation experiment without '$model_{id}$' attribute fed in.

The final ablation results are shown in Figure 31. We can observe that the missing '$price$' feature decreases the accuracy the most and thus used cars' resale price and accidents have a higher bounded relationship than other features.

3.6 Gaussian Mixture Model for Make's Sensitivity towards Accident Records Analysis



(a) (b) (c)

Figure 32: Gaussian Mixture Model Visualization

To investigate the impact differences in accident records on the used cars' price between different makes, we select two makes Hyundai and Cadillac as the representatives for cheap car make and expensive car makes. We conduct the unsupervised clustering using the Gaussian Mixture Model only evaluating two brands' no-accident cars but with different learnt models:

- (1) First, we will learn the Gaussian Mixture Model on combined 'Hyundai and Cadillac no accident cars', with attributes '$mileage$', '$price$', and '$year$'.The clustering accuracy is 0.6262975778546713 with the visualization in Figure 32a.
- (2) Second, we will learn the Gaussian Mixture Model on combined 'Hyundai all cars and Cadillac no accident cars', with attributes '$mileage$', '$price$', '$year$'. The clustering accuracy is 0.6505190311418685 the visualization in Figure 32b.
- (3) Third, we will learn the Gaussian Mixture Model on combined 'Hyundai no accident cars and Cadillac all cars', with attributes '$mileage$', '$price$', '$year$'. The clustering accuracy is 0.6038062283737025 the visualization in Figure 32c.

Analysis: Compared with (1), (2) increases 2.42%. It is because cars with accidents usually have lower prices and Hyundai is the cheaper brand compared with Cadillac. Thus, adding Hyundai cars with accidents into learning will enhance the clustering performance compared with (1). Compared with (1), (3) decreases 2.25%. It is because Cadillac is the more expensive brand compared with Hyundai and adding lower price Cadillac samples (with accident) will make the clustering boundary ambiguous. 2.42% is a little larger than 2.25% which may indicate that the impact of accidents on Hyundai is larger than the impact of accidents on Cadillac. But it is not enough. Considering the number of sample contrasts between 'Cadillac data with accident' and 'Hyundai data with accident': 299 vs. 120, but 'Cadillac data with accident' affects the Gaussian Mixture Model's performance less. We can thus conclude that Cadillac cars' price is more robust against the accident records and Hyundai cars' price is sensitive towards the accident although it is a cheap used car brand. We can approximate the accident record affecting Hyundai used cars' prices 2.68 times more than the influence on Cadillac used cars' prices by computing (2.42/120)/(2.25/299).

4 Real-time Implementation - An Interactive Platform for Resale Price Prediction

4.1 Data Persistence using MySQL

Data persistence is a crucial component of a robust big data machine learning tool. It ensures the longevity of data beyond the execution of the program that created it, enabling smoother refinement of the model as the dataset grows over time.

For our project, a relational database was deemed the most suitable choice due to the structured nature of the dataset obtained from the customized scraping tool. MySQL, deployed on AWS RDS, was selected as the database solution because of the convenience offered by cloud-based systems, including enhanced team collaboration, support, and security provided by a well-established platform. [7]

Specifically, A database named cse587carpredictor was created, with a table called used_cars to store the data, which was initially stored in a simple CSV file. The dataset was preprocessed to meet MySQL's constraints and formatting requirements. Additionally, the pymysql library, a pure-Python MySQL client, was utilized to interface with the database programmatically via Python, facilitating seamless integration between the database and the machine learning workflows.[6]

4.2 Interface for CRUD operation on Database

To allow users with specific authorization to access the database is a critical aspect of a practical data application. Proper authorization ensures secure database maintenance and facilitates efficient management, inspection, troubleshooting, and expansion. For this project, a modern frontend interface was developed using Streamlit, enabling users to perform CRUD (Create, Read, Update, Delete) operations on the dataset in a user-friendly manner. This provides an alternative method to manage and oversee the dataset, complementing direct database access and scraping processes.

A simple yet robust authentication functionality was implemented using Streamlit-Authenticator, a third-party Streamlit library. Users with admin authorization can log in to the admin page by entering their username and password, as shown in Figure 33. Additionally, the admin page incorporates cookie settings, ensuring that users do not need to log in repeatedly when accessing the page.

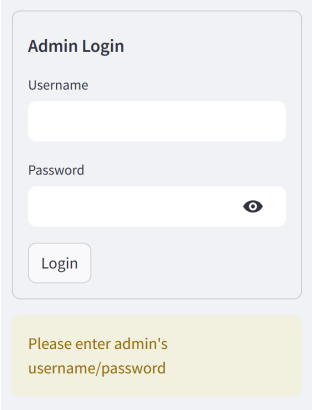
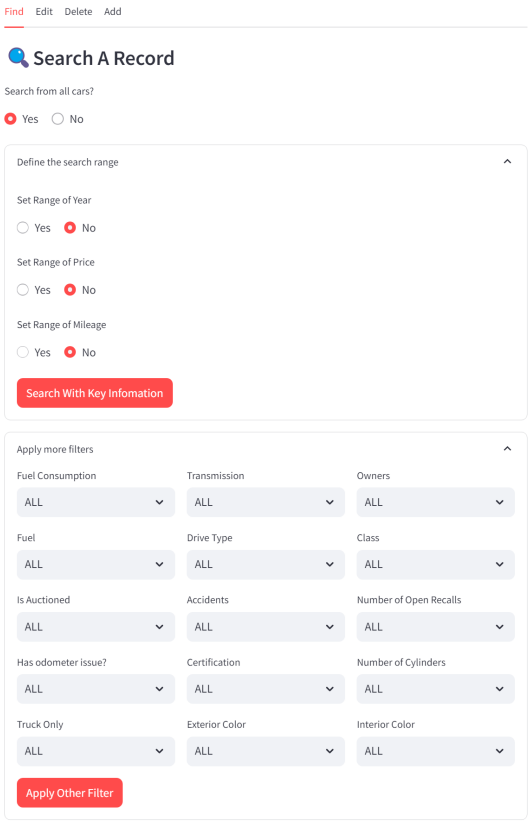
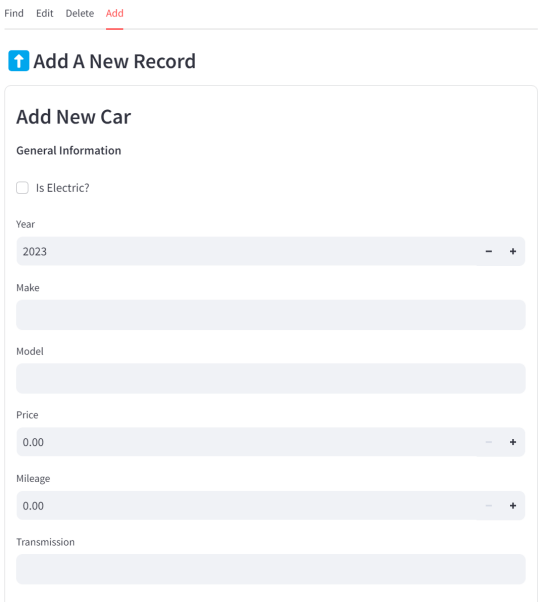


Figure 33: Admin Login

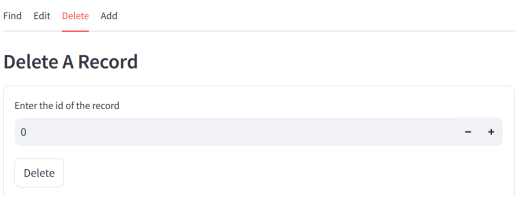
After login, the user can perform CRUD operations on the 4 seperate tabs respectively as shown in refadmin page



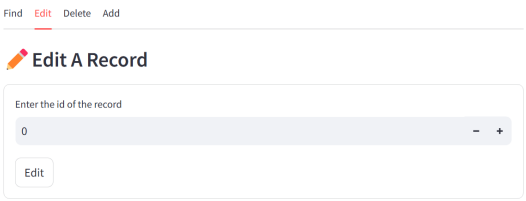
(a) Search Page Tab



(b) Part of Add Record Tab



(c) Delete Record Tab



(d) Edit Record Tab

Figure 34: Tabs: Search, Add, Delete, and Edit Record

Each operation is designed to be straightforward, user-friendly, and efficient. Specifically, the search page offers a variety of search options, enabling administrators to retrieve information as needed, even with the complexity of the dataset. The search functionality incorporates a pre-search dynamic refreshing mechanism. This means that before the SQL query is sent to the database to fetch the required data, unavailable options are filtered out based on the current selections. For example, if a user selects "Honda" for the make, only Honda models will appear in the model dropdown menu. This approach ensures an effective and error-reduced search process.

The delete and edit functionalities are based on the SQL ID of each record. These operations are performed by entering the ID of the car that the user wants to delete or update. For adding new records, the admin has the flexibility to input car information tailored to the specific requirements of the situation.

All CRUD operations include validation checks to ensure the values entered by the user are permissible. For example, prices cannot be negative, and null values are only accepted where appropriate. Additionally, the admin page, like all modern applications, provides meaningful and straightforward prompts to notify the operator whether an operation was successfully performed or not.

4.3 Model Page

In addition to the primary interactive interface, each sidebar integrates distinct models that enable users to input the features of their used car and obtain predictive outcomes. Furthermore, three functional pages are included. The brand analysis page allows users to identify the most competitive brands for their vehicle in the current used car market. The regression model performance page provides users with an interactive experience of the model training process and facilitate an in-depth analysis of the correlations between features and pricing. Lastly, the accident history prediction page enables users to predict a car's accident history based on other features.

4.3.1 Regression Model Performance Page

The objective of the regression model performance page, as is shown in Figure 35 is to evaluate and compare the performance of the CatBoost and Random Forest regression models using metrics such as Root Mean Square Error (RMSE), R^2, and accuracy. Here, accuracy is defined as the proportion of predictions within a 10% deviation from the actual price. To determine the model best suited for used car price prediction, these evaluation metrics are applied to assess performance using all available features.

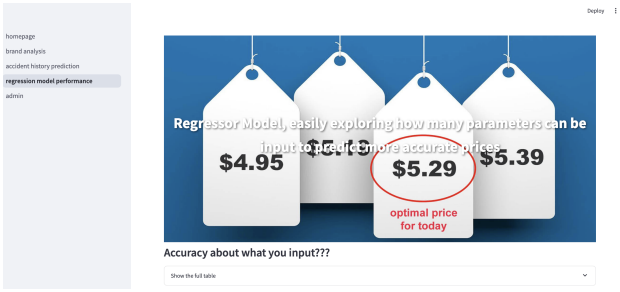


Figure 35: Regression Model Performance Page

Initially, this process is computationally intensive, as it requires training the CatBoost and Random Forest regression models on the complete set of features. Random Forest regression not only predicts prices but also identifies the most important features influencing price prediction.

The results in Figure 36 indicate that the CatBoost regression model outperforms the Random Forest regression model in this context.

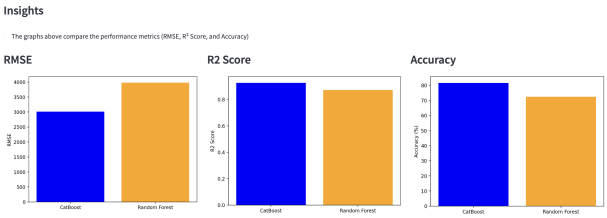


Figure 36: Regression Models Evaluation

As highlighted earlier, not all features are essential for price prediction, and some may act as noise or outliers, negatively impacting model performance. To investigate the influence of features on price prediction, a feature selection tool is provided in Figure 37, allowing users to test subsets of up to 20 top features. The parameter k represents the number of top k important features selected for testing. Accuracy, computed as the percentage of predictions within a 10% deviation from the actual price, is dynamically displayed. Results based on all features are presented as a reference to facilitate comparison with the feature-subset-based evaluations.

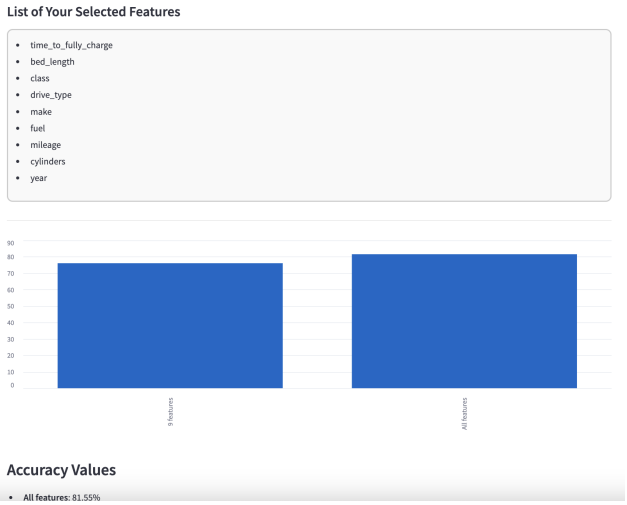


Figure 37: The Impact of Number of Features on Accuracy

As what we can learn from previous content. Not all features are vital features for price prediction and some features even act as noise or outliers which make a worse result. Therefore, to insight the features impact on price, a selected box up to 20 features is provided to test k. k represents the best number of top k important features. Accuracy from computing (accepting within 10% difference) will display dynamically and the results calculated according to all features are shown on the left as a reference.

4.3.2 Brand Analysis Page

In brand analysis, we applied XGBoost to select the top influential brands for each target brand. The importance of using XGBoost has been illustrated in section 3.2. Besides giving an overview table of brands relations, we also provide a snippet for user to explore individual brand influence.

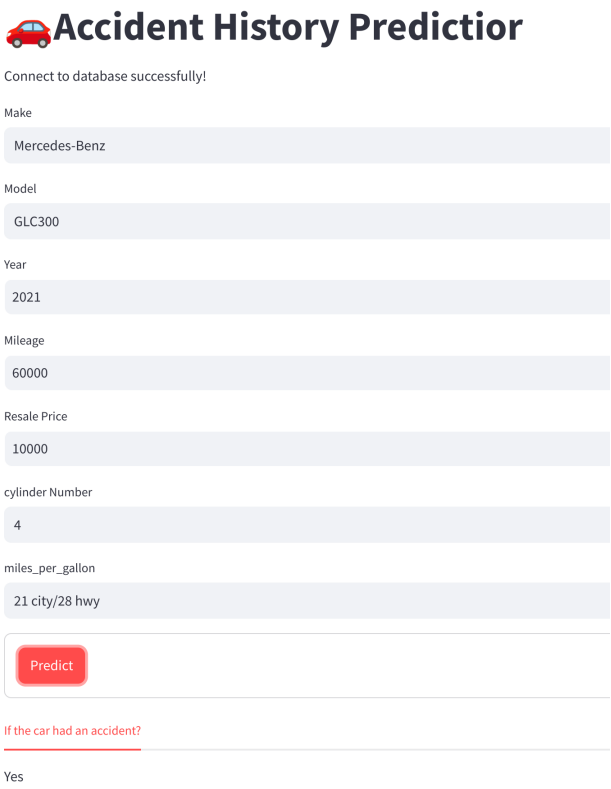
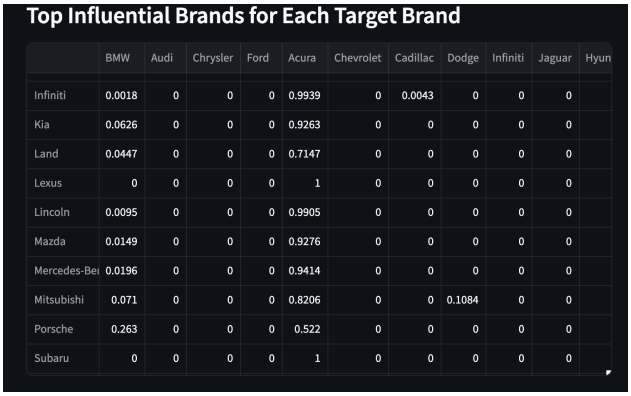


Figure 39: Overview of the accident predictor page



	BMW	Audi	Chrysler	Ford	Acura	Chevrolet	Cadillac	Dodge	Infiniti	Jaguar	Hyun
Infiniti	0.0018	0	0	0	0.9939	0	0.0043	0	0	0	
Kia	0.0626	0	0	0	0.9263	0	0	0	0	0	
Land	0.0447	0	0	0	0.7147	0	0	0	0	0	
Lexus	0	0	0	0	1	0	0	0	0	0	
Lincoln	0.0095	0	0	0	0.9905	0	0	0	0	0	
Mazda	0.0149	0	0	0	0.9276	0	0	0	0	0	
Mercedes-Ber	0.0196	0	0	0	0.9414	0	0	0	0	0	
Mitsubishi	0.071	0	0	0	0.8206	0	0	0.1084	0	0	
Porsche	0.263	0	0	0	0.522	0	0	0	0	0	
Subaru	0	0	0	0	1	0	0	0	0	0	

Figure 38: Overview Spreadsheet for Brand Analysis

As shown in Figure 41, when users select the target brand Acura, they top 3 most influential brands for Acura will be provided.

4.3.3 Accident History Predictor

We deploy the neural network for accident prediction in Section 3.5 here as shown in Figure 39, where you can input the make, model, year, mileage, resale price, cylinder number, and MPG of your car into the page and get the yes or no prediction.

Besides, we also visualize the accident prediction accuracy with different features omitted from your input car brand to help you investigate which feature correlates with the accident records. The visualization shown in Figure 40 reports the gap between each model variant with the base model(no feature omitted).

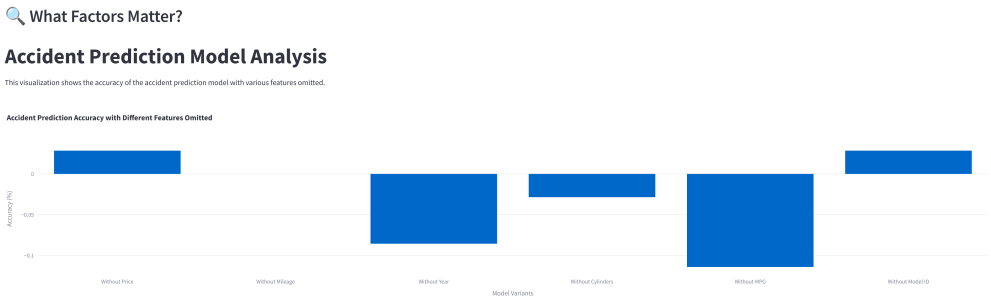


Figure 40: Overview of the accident predictor analysis

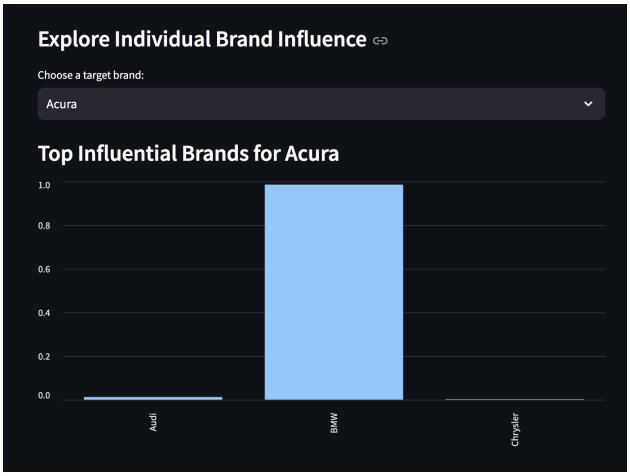


Figure 41: Explore Individual Brand Influence

5 Conclusion

This study explored the factors influencing used car resale prices and developed predictive models using a combination of advanced machine learning techniques. By leveraging models such as KNN, XGBoost, Random Forest, LSTM, CatBoost, and Neural Networks, alongside methods like SMOTENC for data augmentation and Gaussian Mixture Models for clustering, we achieved robust predictions and valuable insights. Key findings include the significant impact of factors like mileage, ownership history, brand, and accident records on resale prices, with brand-specific nuances in sensitivity.

The integration of ensemble and stacking models improved predictive accuracy, while SMOTENC effectively addressed imbalances in underrepresented data groups. Despite these achievements, challenges such as data scarcity for rare categories and limited interpretability in complex models remain. Future work will focus on expanding datasets, incorporating external factors like economic trends, and developing hybrid models that balance accuracy with explainability. This research contributes to a deeper understanding of the used car market and provides a foundation for transparent, data-driven decision-making tools for buyers and sellers.

Acknowledgments

This work was supported in part by the personal node on Center of Computational Research(CCR) provided by Dr. Kang Sun, which facilitated the experiments conducted in this study.

References

[1] S. Pudaruth, "Predicting the price of used cars using machine learning techniques," *International Journal of Information and Computation Technology*, vol. 4, no. 7, pp. 753–764, 2014. [Online]. Available: `https://`

example.com/used-cars-paper

[2] M. Asghar, A. A. Khan, S. Iqbal, and A. Javed, "Used cars price prediction using machine learning with optimal features," *Pakistan Journal of Engineering and Technology*, vol. 4, no. 2, pp. 113–119, 2021.

[3] A. S. Pillai, "A deep learning approach for used car price prediction," *Journal of Science & Technology*, vol. 3, no. 3, pp. 31–50, 2022.

[4] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed. O'Reilly Media, 2017. [Online]. Available: http://cds.cern.ch/record/2699693

[5] Scikit-learn, "GridSearchCV: Exhaustive search over specified parameter values for an estimator," [Online]. Available: https://scikit-learn.org/dev/modules/generated/sklearn.model_selection.GridSearchCV.html. [Accessed: Dec. 4, 2024].

[6] PyMySQL, "PyMySQL Documentation," [Online]. Available: https://pymysql.readthedocs.io/en/latest/. [Accessed: Dec. 4, 2024].

[7] Amazon Web Services, "Amazon RDS for MySQL," [Online]. Available: https://aws.amazon.com/rds/mysql/. [Accessed: Dec. 4, 2024].

[8] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd International Conference on Knowledge Discovery and Data Mining*, 2016.

[9] T. Chen and C. Guestrin, "XGBoost Documentation," 2024. [Online]. Available: https://xgboost.readthedocs.io/en/latest/

[10] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[11] PyTorch, "torch.nn.LSTM," 2024. [Online]. Available: https://pytorch.org/docs/stable/generated/torch.nn.LSTM.html

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[13] Imbalanced-learn, "SMOTE: Synthetic Minority Over-sampling Technique," 2024. [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.over_sampling.SMOTE.html

[14] L. Breiman, "Stacked regressions," *Machine Learning*, vol. 24, pp. 49–64, 1996.

[15] Ugoni, Antony, and Bruce F. Walker. "The Chi square test: an introduction." COMSIG review 4.3 (1995): 61.

[16] SciPy Community, "Scipy Documentation," 2024. [Online]. Available: https://docs.scipy.org/doc/scipy/reference/index.html

[17] CatBoost Developers, "CatBoost Documentation," 2024. [Online]. Available: https://catboost.ai/docs/

[18] Scikit-learn Developers, "Scikit Documentation," 2024. [Online]. Available: https://scikit-learn.org/stable/