Logistic Regression of Toyota Used Car Market Insights

Tyler Klink, tklink@bellarmine.edu

This project investigates whether a used Toyota vehicle is priced above or below the median market price based on its attributes. Using a dataset of more than 6,700 Toyota listings obtained from Kaggle, I created a binary target variable, High Price, where 1 indicates a vehicle priced above the dataset's median price.

Understanding what drives vehicle pricing is meaningful both academically and practically. Buyers often rely on features such as mileage, age, engine size, or transmission type to assess whether a price is fair. By applying logistic regression and a complete machine learning pipeline, this project demonstrates how well vehicle features can classify a car as high-priced and which attributes have the greatest influence on price.

The dataset contains 6738 rows and 9 original columns, representing features of Toyota used cars. These include:

Price	The listed Selling Price (Numerical - continuous)
Year	Manufacturing Year (Numerical - continuous)
Mileage	Distance Driven (Numerical - continuous)
Fuel Type	Petrol, Diesel, Hybrid, etc. (Categorical)
Transmission	Manual, Automatic, Semi-Auto (Categorical)
Engine Size	Engine Displacement (Numerical - continuous)

MPG	Fuel Efficiency (Numerical - continuous)
Tax	Annual Vehicle Tax (Numerical - continuous)
Model	Model Name (Categorical)

Two new variables were engineered:

Age	Newest_year – car_year (Numerical – Continuous)
High Price	1 if Price > median price; 0 otherwise (Categorical – Binary)

The dataset comes from Kaggle, a widely used, reputable open-data platform frequently used for machine learning projects. This makes it a credible and appropriate dataset for classification tasks.

Exploratory data analysis revealed several clear patterns related to vehicle pricing. Boxplots showed that cars in the high-price category generally have much lower mileage than those priced below the median (Figure 1). Scatterplots demonstrated a strong negative relationship between vehicle age and price, with newer vehicles more likely to be high-priced (Figure 2). Tableau visualizations further indicated that automatic and semi-automatic transmissions (Figure 3), as well as hybrid fuel types Figure 4), appear more frequently among higher-priced vehicles.

The High Price target variable was relatively balanced after binarization, making it suitable for logistic regression (Figure 5). Correlation analysis showed that mileage and age are

negatively correlated with price, while engine size is positively correlated. Although some outliers were present, such as very high-mileage or unusually expensive vehicles, they did not significantly affect the analysis. Only a small number of observations contained missing values, which were removed without meaningfully reducing the dataset.

Several preprocessing steps were applied prior to modeling. Rows with missing values in essential features were removed to ensure data quality. Categorical variables, including fuel type, transmission, and model, were encoded using one-hot encoding so they could be used in the logistic regression model. Numerical variables such as mileage, tax, mpg, engine size, and vehicle age were standardized to place them on a comparable scale. The data was then split into training and testing sets using a 75/25 stratified split to preserve the distribution of the High Price classes.

A logistic regression model was implemented using a scikit-learn Pipeline that combined preprocessing and model training into a single workflow. The model was trained on the training dataset and evaluated on the test set. Performance metrics indicated strong results, with overall accuracy between approximately 87% and 90% and balanced precision, recall, and F1-scores across both classes. The confusion matrix showed that the model consistently classified both high-price and low-price vehicles correctly.

Coefficient analysis provided insight into the drivers of pricing. Higher mileage and greater vehicle age reduced the likelihood of being classified as high-priced, while larger engine size increased it. Automatic and semi-automatic transmissions, as well as hybrid fuel types, were also associated with higher prices.

The results demonstrate that vehicle characteristics can effectively predict whether a Toyota is priced above or below the median. Mileage, age, and engine size were the strongest numerical predictors, while transmission type and fuel type added meaningful explanatory power. The model's performance and coefficient signs closely aligned with patterns observed during exploratory analysis and with real-world expectations of the used car market.

This project shows that logistic regression is a reliable and interpretable method for classifying used Toyota vehicles by price category. Newer vehicles with lower mileage, larger engines, automatic transmissions, and hybrid fuel types were more likely to be priced above the median. While the model performed well, it may be limited by the absence of variables such as vehicle condition, trim level, or location. Future work could incorporate these features or explore alternative classification models to further improve performance.

Figure 1:

Mileage by High Price

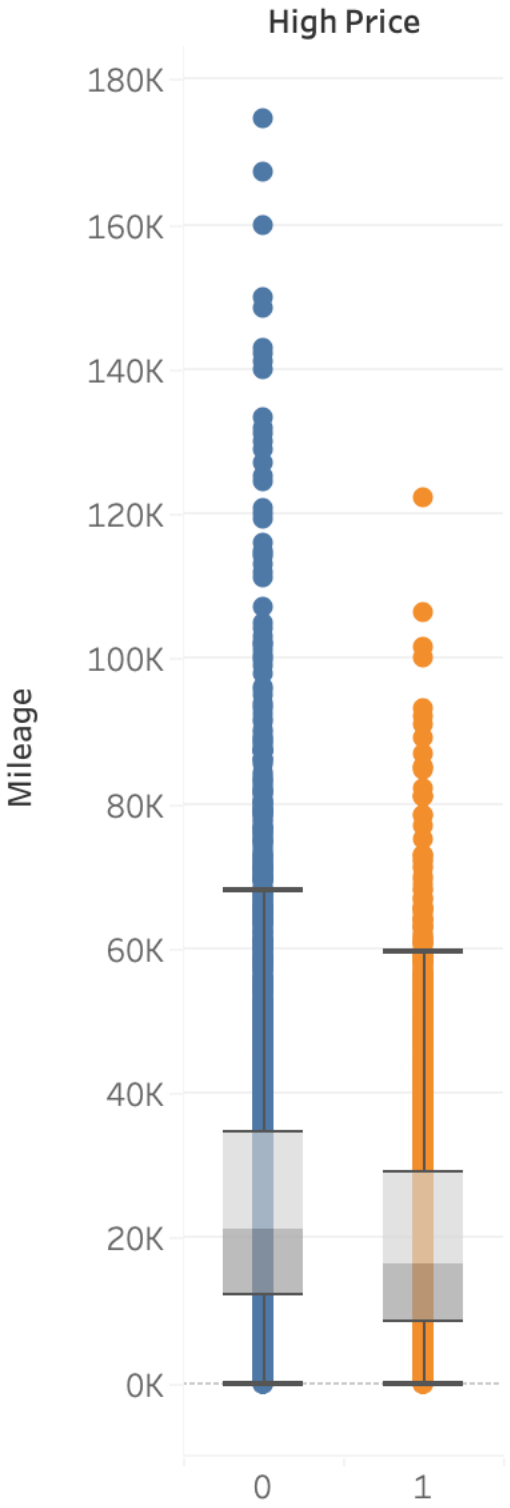

Figure 2:

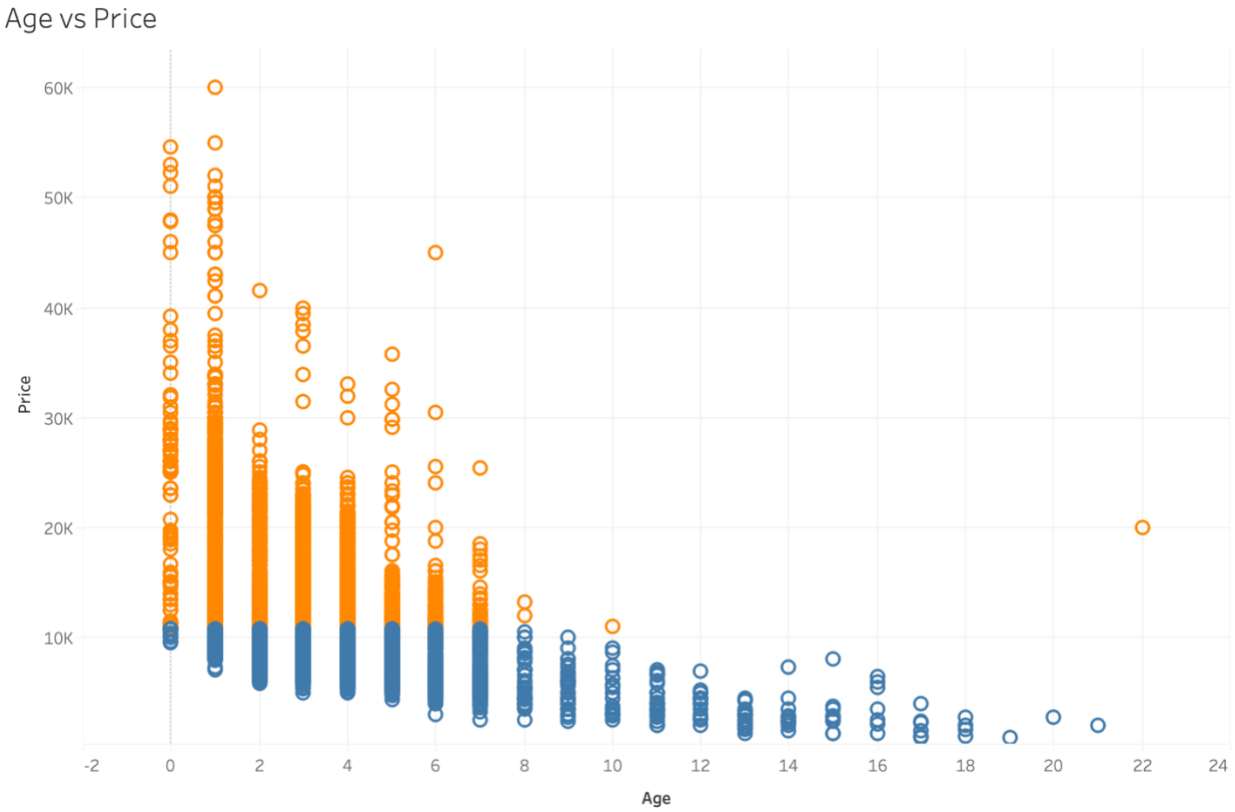

Figure 3:

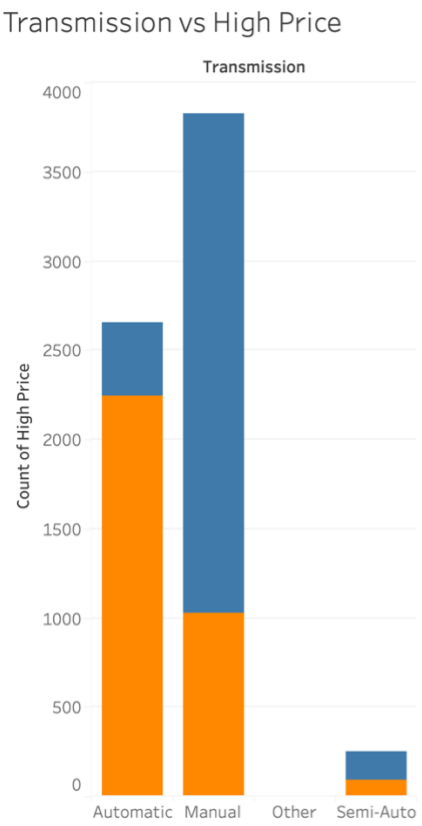
Transmission vs High Price
Transmission
4000
3500
3000
2500
2000
1500
1000
500
0
Count of High Price
Automatic
Manual
Other
Semi-Auto

Figure 4:

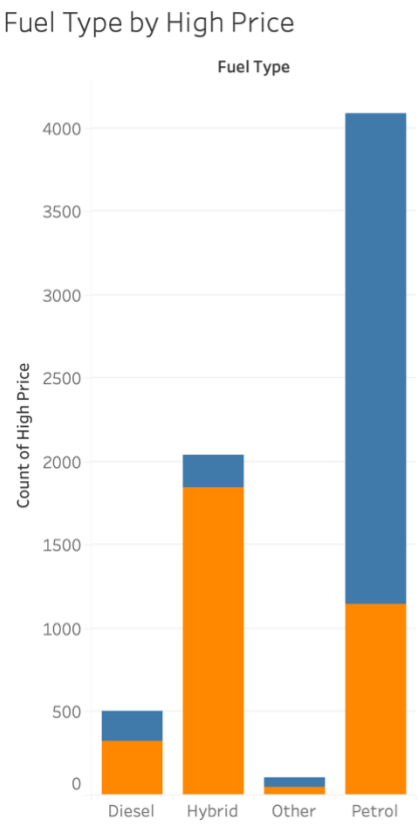
Fuel Type by High Price
Fuel Type
Count of High Price
4000
3500
3000
2500
2000
1500
1000
500
0
Diesel
Hybrid
Other
Petrol

Figure 5:

Class Distribution