MBAI 5100
Business
Analytics
Final Project

Predicting Listing Accuracy and Fair Market Value in Used Cars

Aheli Das, Ejiroghene Akpokiro, Linghui Feng, Sushmita Behera, Vruti Jayesh Tailor

1. INTRODUCTION

The used auto market has always been characterized by information asymmetry. In most cases, the sellers have more knowledge of the actual mechanical condition of the vehicle and its past record than the buyers. The project then aims at assessing the reliability of information regarding used cars and to develop a Trust Score system to improve the accuracy of listings and buyer confidence. To this end, we developed a predictive model that had two main objectives: determining the fair market value (FMV) of the vehicle and assessing the reliability of the vehicle listings.

The project is crucial because it involves monetary risks, which the common consumer has to endure. The value of a vehicle is the second-largest asset most families own. A lack of proper valuation and verification process will place the buyers under huge financial risk because they will either pay huge amounts for vehicles, which are already losing value, or they may be tricked by fraudulent activities. The project allows the consumers to understand the worth of their purchases by informing them of the prevailing market value to enable them to make the informed decisions.

Additionally, this project is essential for enhancing the effectiveness of automotive marketplace activities. Uncertainty is the main cause of friction during transactions and sellers and buyers who experience doubt will choose to postpone the deal or demand excessive inspections or choose not to

transact altogether. The research establishes a better market atmosphere by assessing the level of trustworthiness and providing an open FMV to make transactions smoother.

2. DATASET DESCRIPTION

The data used for this project was sourced from a publicly available Kaggle dataset titled ***"Used Car Listings for US and Canada"***. The dataset contains two separate files, one for United States listings and one for Canadian listings. For this project, the United States dataset was selected due to its larger sample size and broader representation of the used-car market. The dataset carries a Kaggle usability score of 8.82, reflecting its completeness, clarity, and suitability for analytical tasks.

The dataset is derived from MarketCheck's automotive data archive, which aggregates over eight years of vehicle inventory collected across the United States and Canada. MarketCheck continuously crawls more than 65,000 dealership websites each day, producing one of the most comprehensive and up-to-date sources of used-car listing information available. The data includes historical VIN-level records capturing changes in each listing from its earliest online appearance to the most recent update. For the scope of this project, the analysis focused specifically on listings spanning the years 2013 to 2022, ensuring a decade-long view of market trends and vehicle valuation patterns.

The United States dataset includes 21 features, covering both vehicle and seller details: id, vin, price, miles, stock_no, year, make, model, trim, body_type, vehicle_type, drivetrain, transmission, fuel_type, engine_size, engine_block, seller_name, street, city, state, and zip. These features provide extensive information required to assess listing reliability and to estimate fair market values using machine learning techniques. In total, the project utilized ***5,649,563*** raw data points after cleaning, provides a large and varied sample that is meaningful for our analysis. Before conducting modelling and evaluation, the dataset underwent cleaning and preprocessing to ensure data quality, address inconsistencies, and prepare the features for predictive analysis.

3. DATA CLEANING

3.1. Loaded the raw dataset and checked missing values

After loading the original us-dealers-used.csv file, the dataset contained over ***6.5 million*** rows and several columns with large amounts of missing values. For example:

i. price had 656,779 missing values
ii. miles had 67,290 missing values
iii. engine_size had 103,266 missing values
iv. engine_block had 107,110 missing values

3.2. Dropped unnecessary columns

We removed four columns that were not useful for analysis: id, stock_no, seller_name, and street. These columns only contained dealer-specific information and did not affect price, mileage, or any modelling task. After dropping them, the structure of the dataset remained the same except with fewer columns.

3.3. Filtered the dataset to vehicles manufactured between 2013 and 2022

Since our study focuses on the last decade of used-car activity, we kept only vehicles with model years between 2013 and 2022. The dataset was reduced to ***6,208,179*** rows, and all vehicles in this subset were confirmed to fall within the 2013-2022 range (min year = 2013, max year = 2022). This step removed very old and inconsistent listings.

3.4. Checked missing values again after filtering

After filtering by year, missing values decreased slightly, but there were still many incomplete rows.

For example:

i. price still had 467,562 missing values
ii. miles had 44,818 missing values

iii. model had 2,532 missing

iv. engine_size had 61,652 missing

Because these fields are essential, we needed to remove rows with missing values so the dataset would be clean and reliable.

3.5. Dropped all rows with missing values

We applied dropna() to remove any row that had at least one missing value. The dataset was reduced from ***6,208,179*** rows to ***5,649,563*** rows. This means we removed about ***558,616*** incomplete records, leaving us with a fully populated dataset ready for proper analysis.

3.6. Converted data types and cleaned text fields

We corrected the data types for the remaining columns:

i. Converted price and year to numeric

ii. Removed extra spaces from make and model

iii. Cast year to integer to avoid decimal issues

These ensured that the fields could be grouped, filtered, and modelled without errors.

3.7. Created log-transformed variables to handle skewness

The distributions for price and miles were heavily skewed. To stabilize these values, we created: log_price, log_miles. This step did not change the original data but added two new columns to support better performance in modelling and visualization.

3.8. Identified outliers but did not remove them

Using the IQR method for price and miles, we created two flag columns- Price_outlier_flag, miles_outlier_flag. We kept the original values because extreme listings in used-car markets (like luxury cars or fleet vehicles) are often real and meaningful. After all cleaning steps, the final dataset contained:

5,649,563 rows, ***17*** columns, no missing values, standardized numeric and text fields, Log-transformed features, Outlier flags for price and miles.

4. PROPOSED ANALYSIS

4.1. Exploratory Data Analysis (EDA)

i. Boxplots for Numeric Features - We have used boxplots to check the distribution of the numerical variables such as price, mileage, engine size, and model year. Boxplots are ideal to identify the outliers as they need to be cleaned or handled before preparing the models. This helped us ensure that extreme values are not affecting our predictive models.

ii. Correlation Heatmaps Before and After Feature Engineering - We have used correlation heatmaps to assess linear relationships between the variables and detect multicollinearity. We observed strong correlations in the pre-engineering heatmap such as year vs mileage. This led us to feature engineering. In the post-engineering heatmap, we saw improvements from the newly created variables such as vehicle age and miles per year. This confirmed much stronger relationships with price.

iii. Bar Plots for Listings and Mean Mileage by Year - We have used bar plots to analyse the patterns in market behaviour and vehicle usage trends across years. Listing counts by year helped us to identify data distribution over time. For example, we observed larger vehicle concentration during 2017-2019, especially in 2018. Also mean mileage by year revealed expected patterns where with the increase in vehicle year (i.e., newer cars), the mean mileage dropped sharply.

4.2. Linear Regression Models

After splitting our dataset, we prepared 2 linear regression models. Both models were trained on the training set and both of them use log price as the target variable.

i. Model 1 - This uses numerical predictors miles, vehicle age, and engine size, and categorical predictors make, segment, fuel category, transmission, and drivetrain.

ii. Model 2 - This model includes numerical predictors vehicle age, miles per year, and engine size, and categorical predictors engine block, make, segment, fuel category, and transmission, removing drivetrain.

We will assess the models using their RMSE, R^2, and adjusted R^2 to compare predictive accuracy and then further perform Residual analysis to check model assumptions and identify any non-linearity.

4.3. Trust Model

We have developed a logistic regression model to predict a trust flag. Our logistic regression model uses trust flag as target variable and the predictor variables include vehicle age, miles per year, engine size, engine block, segment, fuel category and transmission. The trust score indicates whether a price prediction is trustworthy or not based on segment-wise residual analysis. This will help us quantify the confidence level of each prediction and give us a more transparent interpretation.

4.4. Forecasting

We have applied Holt's linear trend model to forecast the median price of used cars over time. We used median price by year to summarize the central tendency of used-car prices annually and to reduce the impact of outliers and price variability. Holt's model helped us to generate next 3-year forecasts that account for underlying trends in the market without considering seasonal patterns. These forecasts can help dealers anticipate price shifts and make required pricing decisions. Additionally, our model is flexible enough to adjust with upcoming new data and offer a continuous market monitoring.

Technical Notes:

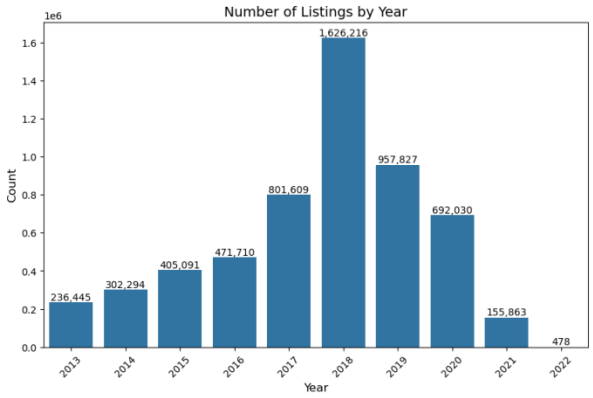
i. We divided our dataset into 80% training set and 20% validation set. This split ensures that we have a robust and more unbiased model evaluation. We fit our models on the training datasets and the validation set was used to assess the models' predictive performance.

ii. We observed quite a few outliers during our analysis. Those were carefully flagged but we chose to retain those as they reflect real world market dynamics where rare or high model cars might have extreme prices. Hence, retaining the outliers ensures that our models do not underestimate pricing variability in the market.

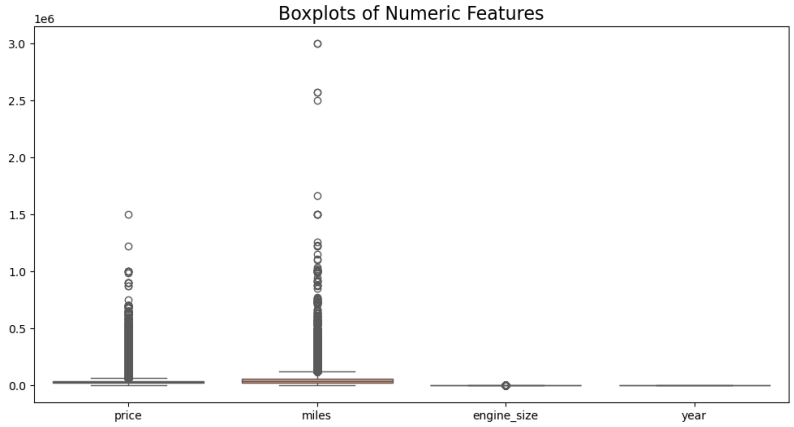
iii. As part of feature engineering, we have created variables such as miles per year, vehicle age, and segment categories to better capture pricing patterns and interactions between factors. We have derived these new features in order to enhance our model's predictive power and accuracy.

5. EXPECTED RESULTS AND MANAGERIAL INSIGHTS

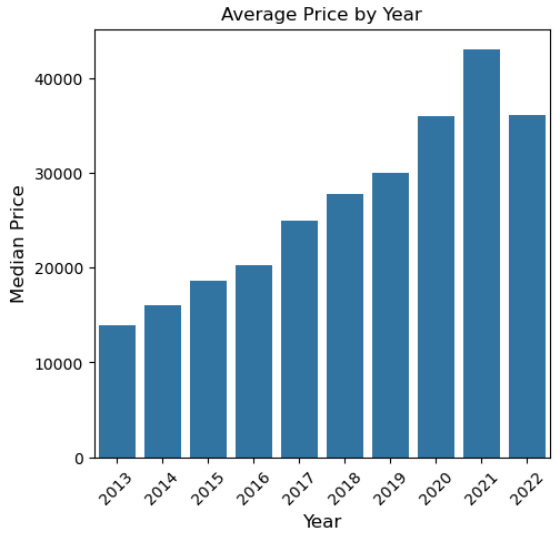
5.1. Number of Listings by Year: The bar chart from shows the distribution of used car listings by year. The highest number of listed car listings is from 2018, followed by 2019 and 2017. We have fewer listings for 2021 and 2022. The reason could be due to limited time on the market or supply chain constraints. Older cars (2013-2015) also have fewer listings, which shows lower market availability and demand.



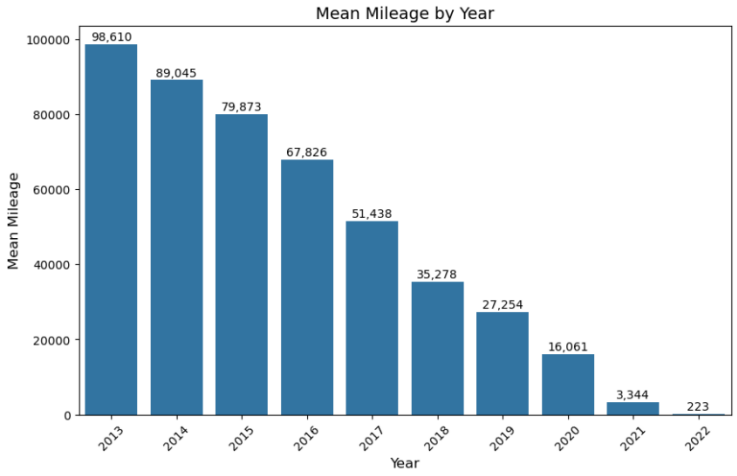
5.2. Boxplots of Numeric Features: The box plot displays the distribution of four numerical features: price, miles, engine_size, and year. Both price and year show many outliers above the upper whisker, indicating the presence of high-priced luxury vehicles and cars with extremely high mileage. Most of the data points are concentrated at the lower values, with a long right tail. Most vehicles have engine sizes clustered around the median, with a few outliers which represent high-performance or special vehicles. The distribution of years is tight, reflecting the earlier filtering for cars manufactured between 2013 and 2022.



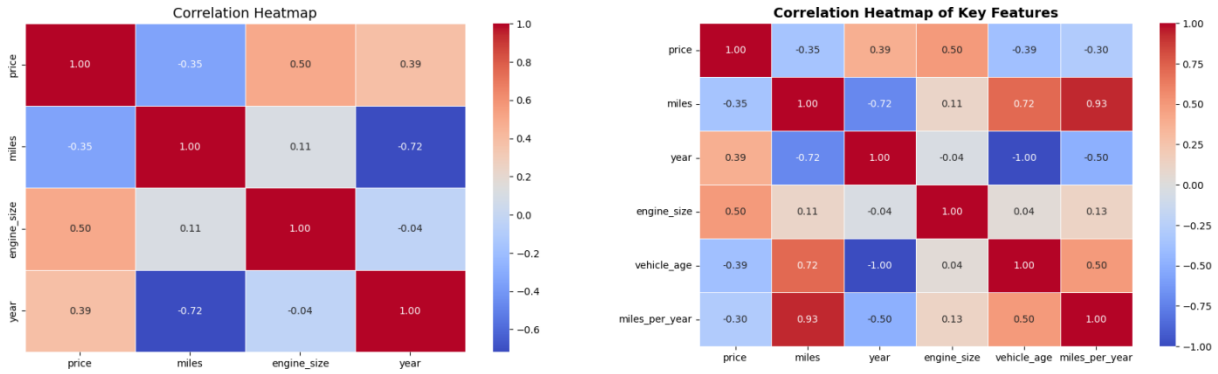
5.3. **Average Price by Year:** The bar chart shows the median prices of used cars by year. There is an upward trend, which means newer vehicles have significantly higher median prices as compared to older models (2013-2016). The highest median price is for the year 2021 models.



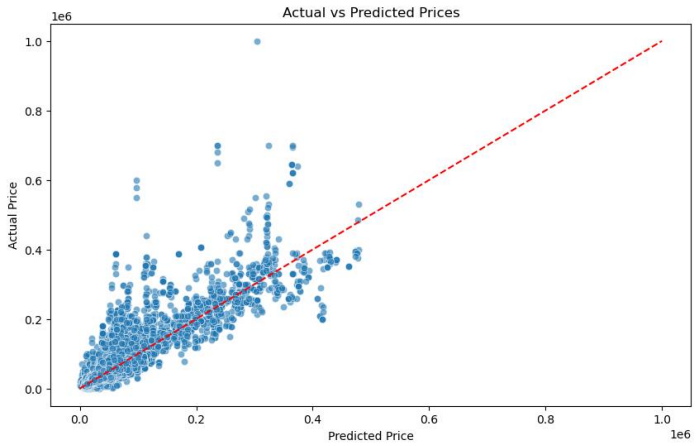
5.4. Mean Mileage by Year: The bar chart shows the average mileage of used cars by year. Older vehicles (2013-2016) have higher mean mileage, while newer vehicles (2020-2022) have lower mean mileage. The pattern is expected as older cars have been driven longer.



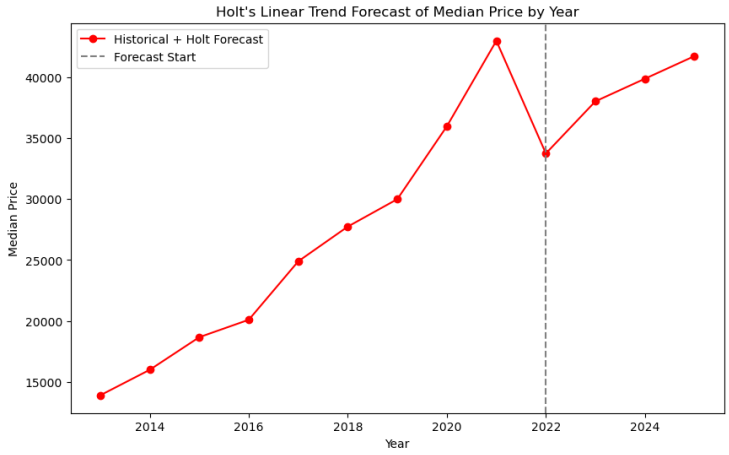
5.5. Correlation Heatmaps (Before and After FE): The correlation heatmaps show that after feature engineering, relationships between the variables are clearer. Price is more strongly negatively correlated with vehicle age and mileage, while miles per year is highly correlated with total miles.



5.6. Actual vs Predicted Price Scatterplots: The scatter plot shows actual vs. predicted prices. Most of the points are close to the red diagonal line, which indicates good model accuracy. We do have some outliers, but overall, predictions align well with actual prices.



5.7. Holt's Linear Trend Forecast: The line chart shows historical and forecasted median used car prices by year using Holt's linear model. Prices move steadily up until 2021, dip in 2022, and increase again through 2025.



5.8. Managerial Insights Based on the Findings:

i. **Key Drivers for the Price:** Vehicle age, mileage, engine size, segment, fuel category, transmission, and drivetrain are the most influential factors affecting used car prices. Segment grouping (e.g., SUV, Sedan, Utility) simplifies pricing and reduces model complexity.

ii. **Market segmentation:** Instead of pricing every make and model separately, grouping cars into broad segments (SUV, Sedan, Utility, etc.) gives you more stable and understandable prices. This helps you set clear pricing strategies for different types of vehicles.

iii. **Trust score for Pricing:** The trust score tells you how reliable a price prediction is. High trust scores mean the price is likely accurate; low scores mean you should double-check before setting a price or negotiating. Use these scores to quickly spot which listings need a closer look.

iv. **Forecasting Price Trends:** Holt's linear trend model shows that recent years' prices are rising, but the trend is gradual. Managers should anticipate moderate price increases and adjust inventory and pricing strategies accordingly.

v. **Handling Outliers:** Outliers in price and mileage are common in used car markets. The model flags these outliers so you can decide if they're luxury vehicles or just unusual cases that need special attention.

vi. **Model Performance:** Multiple regression models were compared, the second model (including miles per year and engine block) performed better. Managers should use models that incorporate usage intensity and technical features for more accurate pricing.

Data Source: https://www.kaggle.com/datasets/tsaustin/us-used-car-sales-data

Code and Dataset: https://drive.google.com/drive/folders/1W0vHUUROqJxeyxjk-6o4SL6-hRaF35Ix?usp=drive_link

APPENDIX

Figure AI: Summary Statistics and Log-Price Distribution

```
# Summary statistics for numeric columns
car_df.describe().T
```

	count	mean	std	min	25%	50%	75%	max
index	5649563.0	3.559881e+06	2.048648e+06	0.0	1.774890e+06	3.551258e+06	5.327816e+06	7.104303e+06
price	5649563.0	2.964123e+04	1.700268e+04	0.0	1.899900e+04	2.599900e+04	3.644000e+04	1.499996e+06
miles	5649563.0	4.441576e+04	3.368609e+04	0.0	2.080300e+04	3.548300e+04	6.123100e+04	3.000000e+06
year	5649563.0	2.017550e+03	1.931474e+00	2013.0	2.016000e+03	2.018000e+03	2.019000e+03	2.022000e+03
engine_size	5649563.0	3.038041e+00	1.330179e+00	0.6	2.000000e+00	2.500000e+00	3.600000e+00	3.000000e+01
log_price	5649563.0	1.017998e+01	4.756409e-01	0.0	9.852194e+00	1.016585e+01	1.050345e+01	1.422097e+01
log_miles	5649563.0	1.030610e+01	1.238652e+00	0.0	9.942901e+00	1.047684e+01	1.102243e+01	1.491412e+01
price_outlier_flag	5649563.0	3.316982e-02	1.790798e-01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00
miles_outlier_flag	5649563.0	3.049174e-02	1.719360e-01	0.0	0.000000e+00	0.000000e+00	0.000000e+00	1.000000e+00

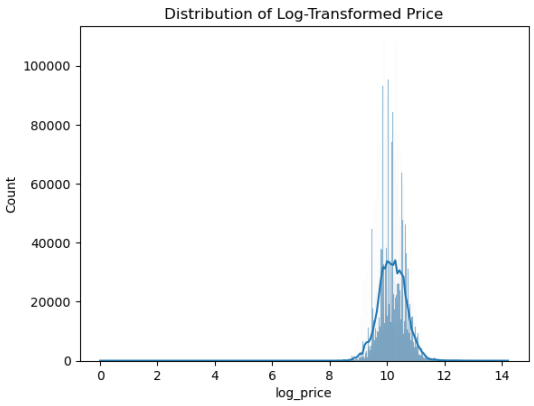
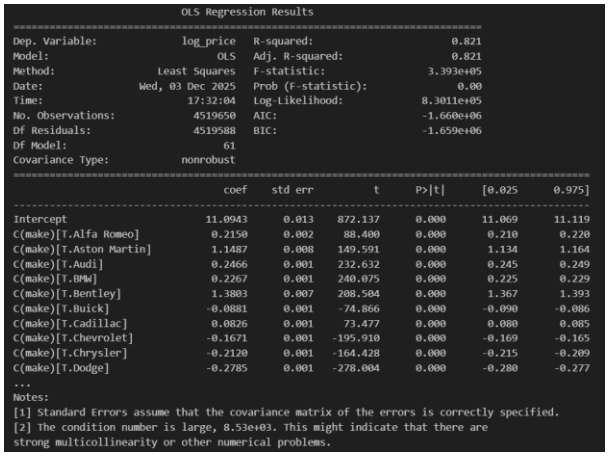


Figure A2. OLS regression results for Model 1 and Model 2

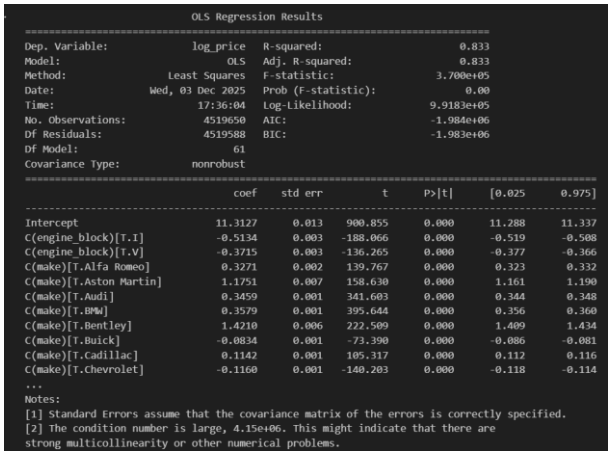


OLS Regression Results

Dep. Variable:	log_price	R-squared:	0.821
Model:	OLS	Adj. R-squared:	0.821
Method:	Least Squares	F-statistic:	3.393e+05
Date:	Wed, 03 Dec 2025	Prob (F-statistic):	0.00
Time:	17:32:04	Log-Likelihood:	8.3011e+05
No. Observations:	4519650	AIC:	-1.660e+06
Df Residuals:	4519588	BIC:	-1.659e+06
Df Model:	61		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.0943	0.013	872.137	0.000	11.069	11.119
C(make)[T.Alfa Romeo]	0.2150	0.002	88.400	0.000	0.210	0.220
C(make)[T.Aston Martin]	1.1487	0.008	149.591	0.000	1.134	1.164
C(make)[T.Audi]	0.2466	0.001	232.632	0.000	0.245	0.249
C(make)[T.BMW]	0.2267	0.001	240.075	0.000	0.225	0.229
C(make)[T.Bentley]	1.3803	0.007	208.504	0.000	1.367	1.393
C(make)[T.Buick]	-0.0881	0.001	-74.866	0.000	-0.090	-0.086
C(make)[T.Cadillac]	0.0826	0.001	73.477	0.000	0.080	0.085
C(make)[T.Chevrolet]	-0.1671	0.001	-195.910	0.000	-0.169	-0.165
C(make)[T.Chrysler]	-0.2120	0.001	-164.428	0.000	-0.215	-0.209
C(make)[T.Dodge]	-0.2785	0.001	-278.004	0.000	-0.280	-0.277
...						

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.53e+03. This might indicate that there are strong multicollinearity or other numerical problems.



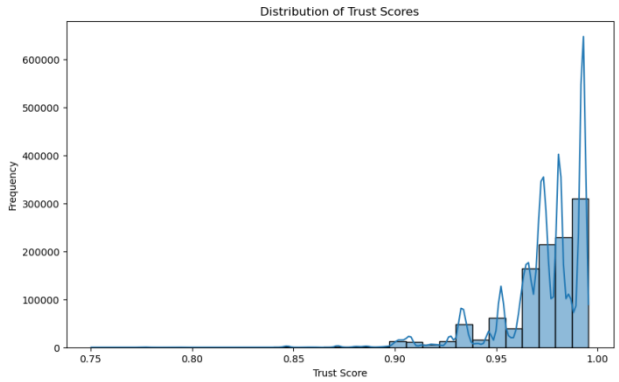
OLS Regression Results

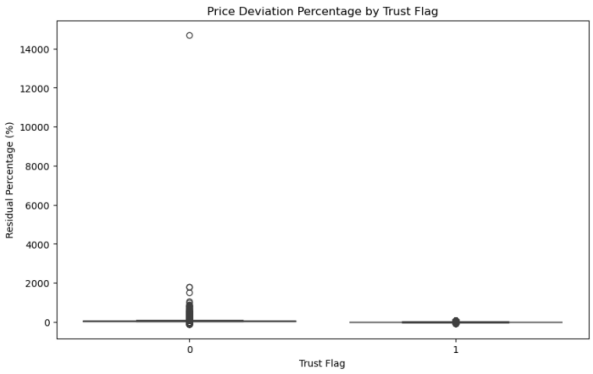
Dep. Variable:	log_price	R-squared:	0.833
Model:	OLS	Adj. R-squared:	0.833
Method:	Least Squares	F-statistic:	3.700e+05
Date:	Wed, 03 Dec 2025	Prob (F-statistic):	0.00
Time:	17:36:04	Log-Likelihood:	9.9183e+05
No. Observations:	4519650	AIC:	-1.984e+06
Df Residuals:	4519588	BIC:	-1.983e+06
Df Model:	61		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	11.3127	0.013	900.855	0.000	11.288	11.337
C(engine_block)[T.I]	-0.5134	0.003	-188.066	0.000	-0.519	-0.508
C(engine_block)[T.V]	-0.3715	0.003	-136.265	0.000	-0.377	-0.366
C(make)[T.Alfa Romeo]	0.3271	0.002	139.767	0.000	0.323	0.332
C(make)[T.Aston Martin]	1.1751	0.007	158.630	0.000	1.161	1.190
C(make)[T.Audi]	0.3459	0.001	341.603	0.000	0.344	0.348
C(make)[T.BMW]	0.3579	0.001	395.644	0.000	0.356	0.360
C(make)[T.Bentley]	1.4210	0.006	222.509	0.000	1.409	1.434
C(make)[T.Buick]	-0.0834	0.001	-73.390	0.000	-0.086	-0.081
C(make)[T.Cadillac]	0.1142	0.001	105.317	0.000	0.112	0.116
C(make)[T.Chevrolet]	-0.1160	0.001	-140.203	0.000	-0.118	-0.114
...						

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 4.15e+06. This might indicate that there are strong multicollinearity or other numerical problems.

Figure A3. Trust score distribution and residual deviation by trust flag





Technical Note

The appendix figures show the key steps that supported our analysis. The summary statistics and log-price distribution confirm that the cleaned dataset has reasonable ranges and that the log transformation produces a more balanced price spread. The two OLS outputs compare our basic model (using only make) with a slightly extended version that also includes engine-block type. The improvement in fit shows that adding mechanical features strengthens the model. The trust-score histogram and the price-deviation plot help illustrate how listings marked as “trusted” behave compared to others. These figures were included to show the checks we ran before finalizing our model.