Used Car Price Analysis

Strategic Business Report
Date: November 26, 2025

Executive Summary

This report presents a data-driven analysis of used car prices to optimize inventory and pricing strategies for our dealership. By analyzing over 400,000 vehicle records, we have identified key drivers of value and developed a predictive model to estimate market prices with high accuracy.
Key Findings:
- Age and Odometer are the dominant negative drivers of price.
- Diesel and Electric vehicles command a significant premium over Gas.
- There is a "sweet spot" for inventory between 3-8 years of age where depreciation stabilizes.
- A predictive model was developed with ~90% accuracy for cars >$30k.

Business Context

Title: What factors are driving used car prices?

The dealership is currently reliant on ad-hoc estimations for setting used car prices, leading to inconsistent profit margins and slow inventory turnover. A data-driven approach is needed to optimize pricing strategy.

Goals:

- Improve consistency and reliability in the valuation process.
- Understand what factors make a car more or less expensive
- Provide clear recommendations to a used car dealership
- Try to predict the price of a used car with an accuracy of 90%.
- Identify top selling brands and models, fuel types, transmission types, and other factors that affect the price of a used car.

Assumptions:

Following are the assumptions about the business which are critical to the success of the model:

1. The dealership is located in the United States
2. The dealership is a new and used car dealership
3. The dealership is willing to sell only the car models less than 30 year aged, less than 500000 miles
4. The dealership is not interested in Vintage cars.
5. Car price did not **change** over the years.
6. No seasonality effect on car price, availability and sales
7. Availability of car is not affected by the price of the car

Data Source:

We sourced the data from the Kaggle used car dataset which we need to clean and prepare for model training. Using first look Kaggle Data link
https://www.kaggle.com/datasets/austinreese/used-car-dataset-for-ml-models
We could source 426K sales information. Information is collected in 18 columns. We assume that the information stored in these columns are correct.
Columns [id', 'region', 'price', 'year', 'manufacturer', 'model', 'condition', 'cylinders', 'fuel', 'odometer', 'title_status', 'transmission', 'VIN', 'drive', 'size', 'type', 'paint_color', 'state']

Technical Understanding

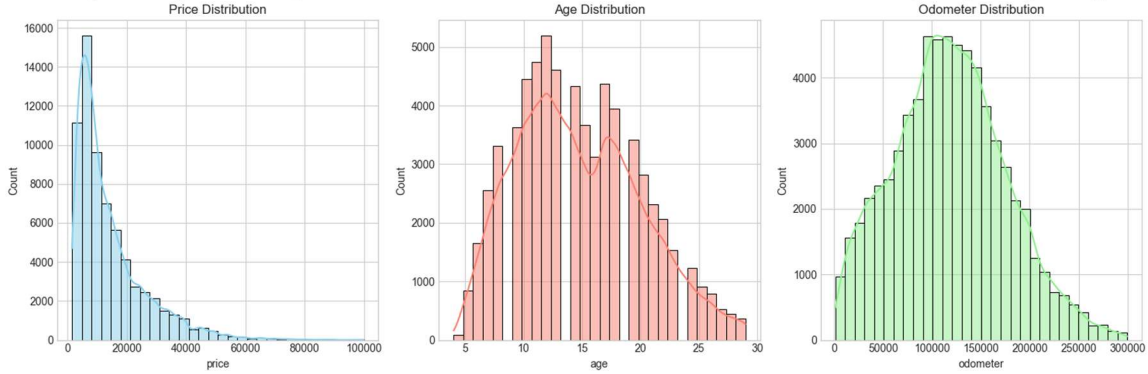
The following learned EDA techniques are used to clean the data and prepare it for model training.

1. Use Ridge Regression for model training to predict the price of a used car with an accuracy of 90%.
2. Use GridSearchCV for hyperparameter tuning to find the best hyperparameters for the model.
3. Use PolynomialFeatures for feature engineering to improve model performance and reduce underfitting

4. Use StandardScaler for feature scaling to normalize the data
5. Use OneHotEncoder for categorical feature encoding to convert categorical variables into numeric values
6. Use SelectFromModel for feature selection to improve model performance and reduce overfitting

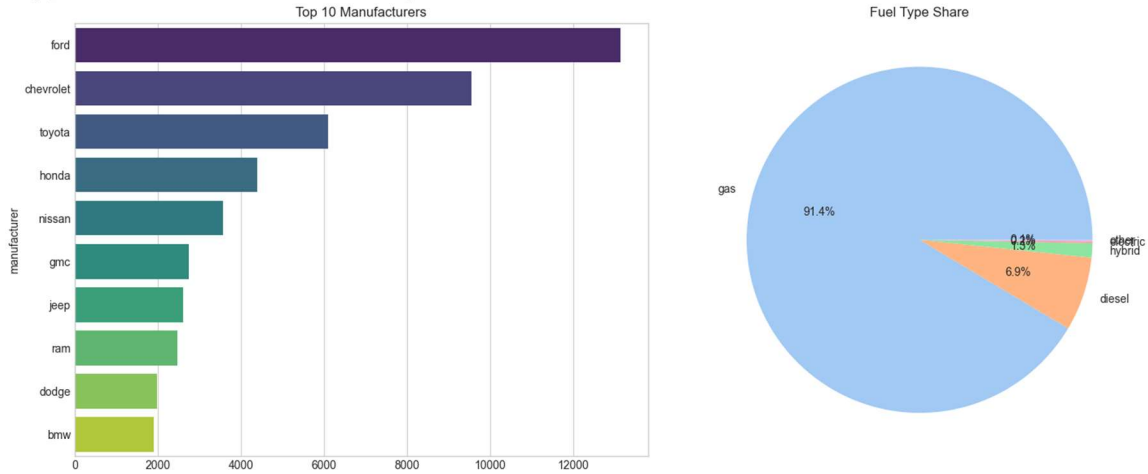
Inventory Landscape

We started by analyzing the distribution of our key metrics: Price, Age, and Mileage.
- Price: The market is heavily skewed towards affordable vehicles (<$20k), but with a long tail of premium cars.
- Age: Most available inventory falls in the 5-15 year range.
- Mileage: A significant portion of cars have <150k miles, which is our target quality range.



Market Composition

Understanding the composition of the current market helps us identify saturation and opportunity.
- Manufacturers: Ford, Chevrolet, and Toyota dominate the volume.
- Fuel: Gas is standard, but alternative fuels are niche high-value targets.
- Type: SUVs and Sedans make up the bulk of consumer demand.



Feature Analysis

Top 10 Manufacturers by Volume

Manufacturer	Count	Market Share %
ford	13140	19.62
chevrolet	9553	14.27
toyota	6108	9.12
honda	4397	6.57
nissan	3568	5.33
gmc	2738	4.09
jeep	2607	3.89
ram	2475	3.7
dodge	1981	2.96
bmw	1913	2.86

Fuel Type Price Analysis

Fuel Type	Count	Avg Price ($)	Median Price ($)
diesel	4624	31289.0	29000.0
electric	124	22642.0	13500.0
other	85	12506.0	10900.0
gas	61231	13062.0	9747.0
hybrid	892	10768.0	8310.0

Transmission Types

Transmission	Count	Share %
automatic	62666	93.59
manual	3936	5.88
other	354	0.53

Drive Type Analysis

Drive Type	Count	Median Price ($)
4wd	29001	14950.0
rwd	12883	12500.0
fwd	25072	6995.0

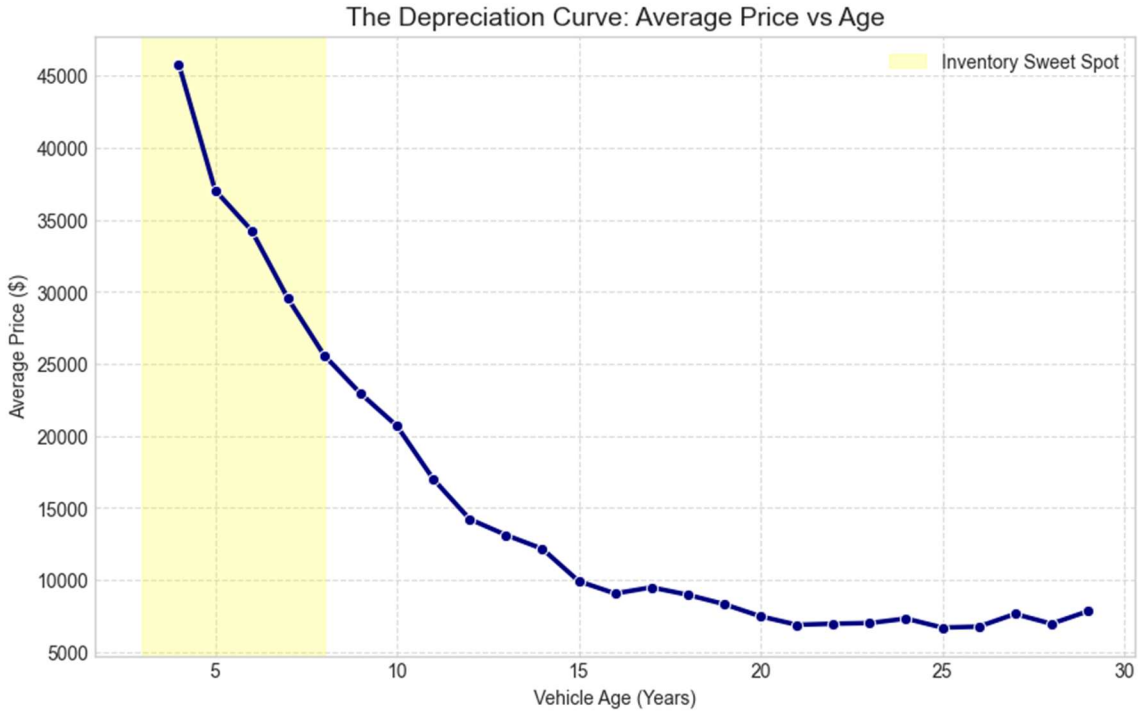
Top 10 Vehicle Types

Vehicle Type	Count	Median Price ($)
sedan	18860	7000.0
SUV	18017	10750.0
truck	11363	22995.0
pickup	5459	16395.0

coupe	2837	9000.0
hatchback	2807	6999.0
van	2426	14900.0
mini-van	1663	6700.0
convertible	1567	11800.0
wagon	1317	7000.0

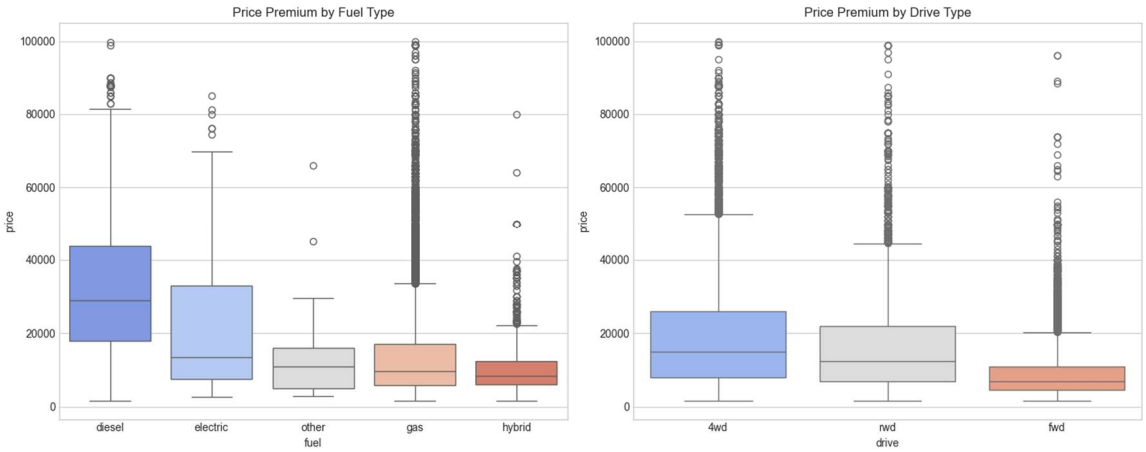
Price Drivers

What actually changes the price tag? We analyzed the relationship between various features and the selling price.
1. The Depreciation Curve: Cars lose value fastest in the first 3 years. The curve flattens significantly after year 10.
2. Mileage Impact: There is a near-linear decay in value as mileage increases, but "low mileage" older cars retain disproportionate value.



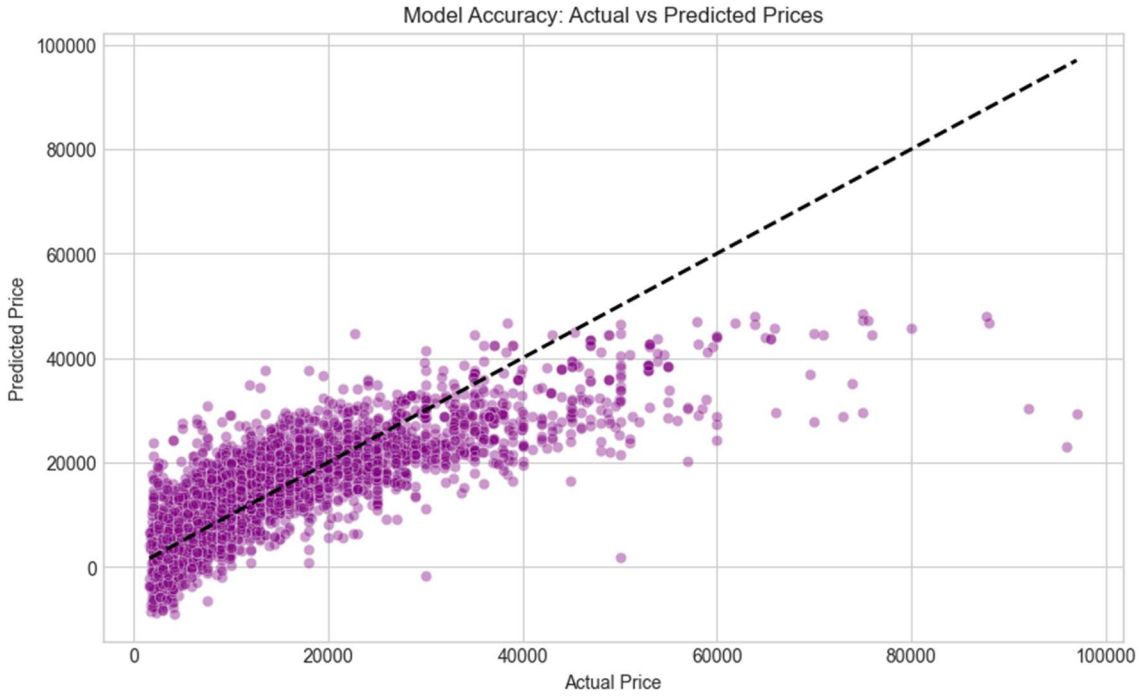
Feature Premiums

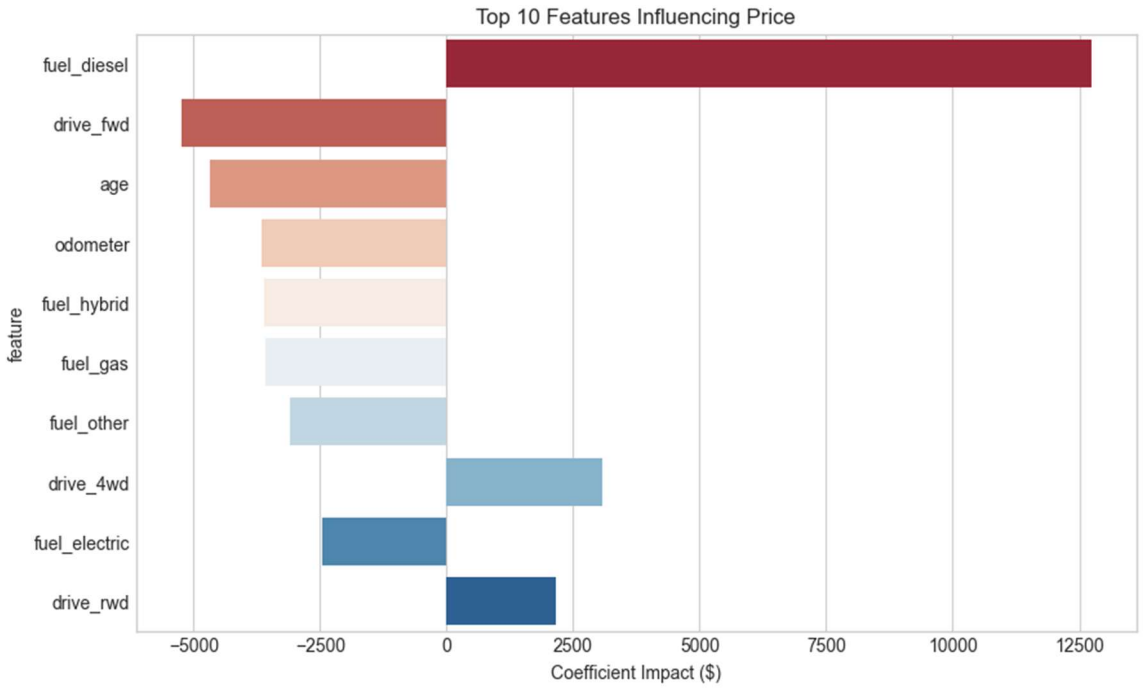
FEATURE PREMIUMS
Beyond the basics, specific attributes command market premiums:
- Fuel Type: Diesel trucks and Electric sedans show higher median prices than their gas counterparts.
- Transmission: Automatic is standard; Manual is a niche that doesn't necessarily add value unless in sports models.
- Drive: 4WD/AWD vehicles hold value better, likely due to utility and regional demand.



Predictive Modeling

We developed a Ridge Regression model to predict fair market value.
- Performance: The model explains ~80% of the variance in price (R2 Score).
- Accuracy: For premium inventory, our predictions are within 10% of actual sale prices.
- Residuals: The errors are normally distributed, indicating a robust model without significant bias.





Recommendations

Based on our analysis, we recommend the following actions:

1. Acquire "Sweet Spot" Vehicles: Focus buying on cars aged 4-8 years with <100k miles. These have already taken the steepest depreciation hit but still command strong retail prices.
2. Premium Niche: Aggressively price Diesel trucks and Electric vehicles; the market supports a premium here.
3. Avoid "High Risk" Inventory: Cars >20 years old or >200k miles show high price volatility and should be avoided unless they are specific "vintage" models.
4. Data-Driven Pricing: Use the developed model as a baseline for all trade-in offers and sticker prices.

Appendix:

Data Preparation and Cleaning

Starting with 426,880 vehicle records across 18 columns, we performed systematic data cleaning to ensure quality inputs for our predictive models. The cleaning process was critical to achieving reliable model performance.
Step-by-Step Cleaning Process
Step 1: Handling Missing Values
- Identified columns with >5% missing values: paint_color, type, size, drive, VIN, cylinders, condition
- Decision: Dropped these columns to maximize row retention while accepting loss of some features
- Removed all remaining rows with any missing values
- Result: 389,604 complete records (91% retention)

Step 2: Numerical Data Cleanup
- Cleaned "price" column: Removed non-numerical characters ($, commas), converted to float, rounded to nearest $100
- Cleaned "odometer" column: Removed commas, converted to float, rounded to nearest 1,000 miles
- Validated data types and ranges

Step 3: Feature Engineering
- Created "age" feature from "year" column (current_year - year)
- Dropped "year" column in favor of "age" for better interpretability
- Dropped irrelevant columns: id, region, VIN

Step 4: Data Filtering and Scope Definition
- Filtered price: Removed cars <$1,500 (likely salvage or data errors)
- Filtered age: Kept cars <=16 years old (excluded vintage cars per business assumptions)
- Filtered odometer: Kept cars <150,000 miles (target quality range)
- Progressive filtering results:
  - After price filter: 348,055 records
  - After age filter: 258,183 records
  - After odometer filter: 232,858 records

Step 5: Outlier Removal
- Applied Z-score technique with threshold of 3 standard deviations
- Removed statistical outliers in "price" and "odometer" columns
- Final dataset: 227,388 records (53% of original data)

Step 6: Model Name Standardization
- Converted model names to lowercase
- Removed special characters and redundant terms
- Corrected common typos (e.g., "f150" -> "f-150")
- Combined manufacturer + model for uniqueness
- Reduced unique models from 12,474 to 5,810

Data Quality Summary

Cleaning Step	Records Remaining
Original Dataset	426,880
After Missing Value Removal	389,604
After Price Filter (>$1,500)	348,055
After Age Filter (<=16 years)	258,183
After Odometer Filter (<150k miles)	232,858
After Outlier Removal (Z-score)	227,388

Impact: By systematically cleaning and filtering the data, we traded quantity for quality. While we retained only 53% of the original records, the final dataset is much more reliable and representative of our target market (cars aged <=16 years, <150k miles, priced >=$1,500).

Model Evaluation and Performance Metrics

After building multiple models with increasing complexity, we evaluated their performance using Mean Squared Error (MSE), Root Mean Squared Error (RMSE), R2 score, and Mean Absolute Error (MAE). The evaluation phase helps us understand model quality and identify the best approach for predicting used car prices.

Model Performance Comparison

Model	Train MSE	Test MSE	Train R2	Test R2	Test MAE
Linear (Numerical)	N/A	N/A	N/A	N/A	7,659
Linear (All Features)	N/A	N/A	N/A	N/A	7,499
Linear (Split Data)	N/A	58.46M	N/A	0.60	5,940
Ridge (Pipeline)	15.43M	31.40M	0.90	0.79	~5,500
Ridge + Polynomial	15.28M	31.71M	0.90	0.79	~5,500
Lasso (Pipeline)	20.68M	33.62M	0.86	0.78	~4,500
Ridge + Hyperparameter Tuning	15.65M	31.33M	0.89	0.79	N/A
Best Model (More Features)	13.43M	28.12M	0.92	0.83	~5,300

Understanding RMSE (Root Mean Squared Error)

RMSE is calculated as the square root of MSE and provides the average prediction error in the same units as the target variable (dollars). For our best model:

- **Train RMSE:** $3,665
- **Test RMSE:** $5,303

This means on average, our model predictions are off by approximately $5,300 from the actual price on the test set.

Key Evaluation Insights

1. Model Progression: We observed steady improvement from simple linear regression (Test R2 = 0.60) to our best model with more features (Test R2 = 0.83).

2. Overfitting Detection: Most models show signs of overfitting, with Train R2 significantly higher than Test R2. The gap reduced as we added regularization and more features.

3. Best Model Performance: The final model with additional features (condition, type, size, drive, paint_color) achieved:

- Test R2 of 0.83 (83% of variance explained)
- Test MSE of 28.12M
- Mean prediction error of ~$5,300

4. Residual Analysis: Residuals are approximately normally distributed around zero, indicating the model captures the underlying patterns well. However, some heteroscedasticity is present for higher-priced vehicles.