# Let Us Sell Your Car

## Let Us Sell Your Car: Leveraging Machine-Learning & CarGuru Data to Predict Vehicle Prices

By: Jessica Gallardo, Shuteng Ong, and Chloe Roque

In the United States, **the used car market plays a significant role** in **the automotive industry, offering consumers** more **affordable alternatives** to new vehicles. **With** rising **inflation, changing consumer preferences, and supply chain disruptions** affecting new car production, **the demand** for used cars **has surged** in recent years. Understanding what drives these price changes is important for sellers navigating this competitive market.

To explore these trends, we used a CarGuru 2021 Inventory dataset from Kaggle, which includes 217,000 rows and 60 columns of detailed car listings. **This dataset** contains **features** such as **make and model, age, mileage, condition, vehicle history, and much more,** offering a comprehensive view of the used car landscape.

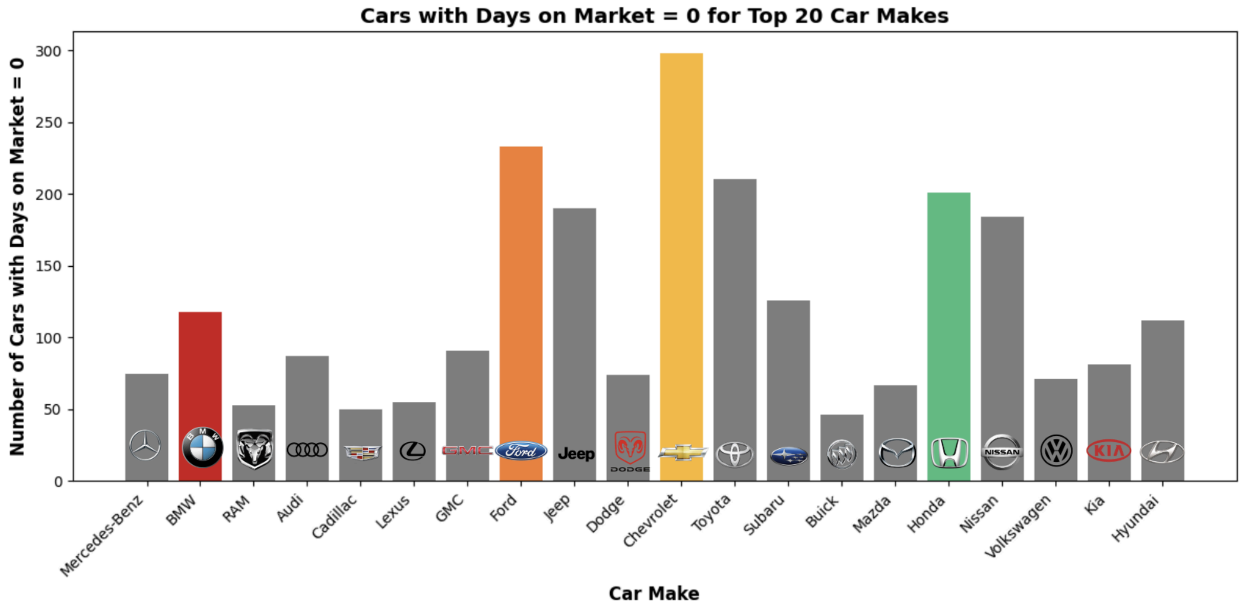
**Our research** focuses on the **question: What features influence used car prices, and how many days are cars typically on the market?**

### Data Analysis - Days on Market & Price

To begin our investigation, **we focused on the top 20 car makers,** which **account for approximately 90% of the dataset**, excluding the remaining 33 makers that make up only 10%. We then categorized these top 20 makes into four price categories based on their average resale price:

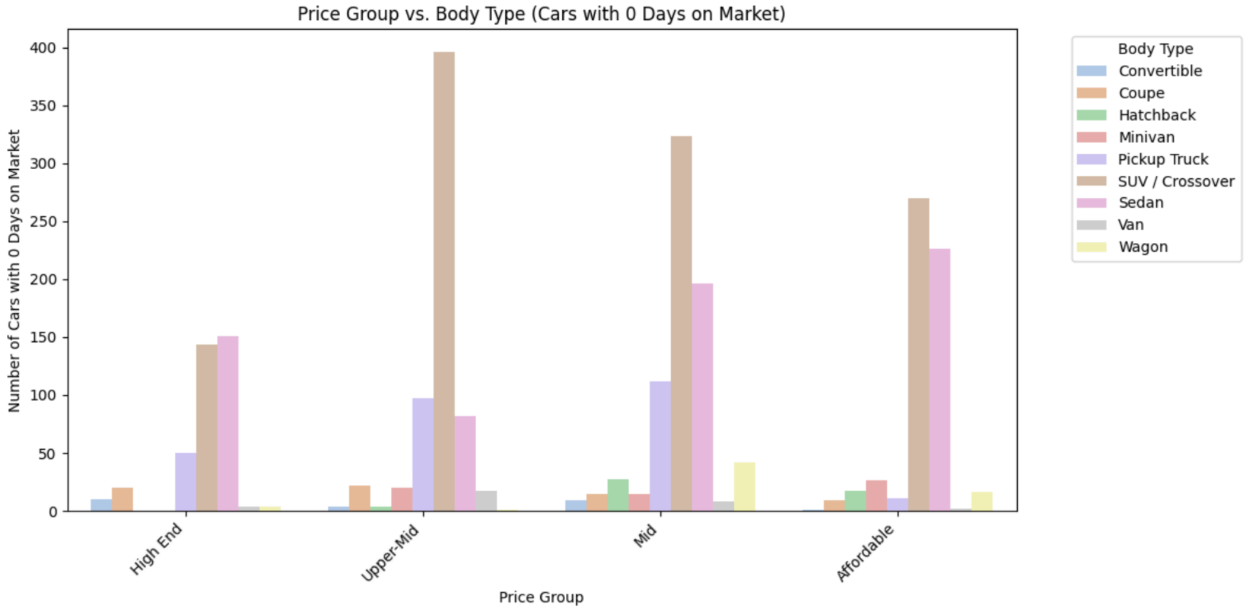
- **High-end range ($35k-$27k):** Mercedes-Benz, BMW, RAM, Audi, Cadillac
- **Upper-mid range ($26k-22k):** Lexus, GMC, Jeep, Dodge, Ford
- **Mid range ($21k-17k):** Chevrolet, Toyota, Subaru, Buick, Mazda
- **Affordable ($17k-16k):** Honda, Nissan, Volkswagen, Kia, Hyundai

Among our four price groups, **we observed that BMW, Ford, Chevrolet, and Honda** cars **had the highest number of listings with 0 days on the market**, **meaning** they were **sold immediately** or on the same day they were listed, as seen below in Figure 1. Figure 1 below displays the top twenty cars in order of average prices, with the car brand logos represented. The red, orange, yellow, and green are the listed cars with the highest number of listings with 0 days on the market, as discussed before.



After exploring which car brands tend to sell immediately, we extended our analysis to examin e which vehicle body types are most commonly listed with 0 days on the market, indicating immediate or same-day sales. To add more context, we broke this analysis down by price group to uncover how body type preferences vary with consumer spending habits.

As shown in Figure 2 below, **Pickup Trucks, SUVs/Crossovers, and Sedans emerged as** the **most in-demand body types** across all price categories: High-End, Upper Mid, Mid, and Affordable.



Our findings indicate that across all price tiers, certain brands and body types consistently dominate when it comes to vehicles listed with zero days on the market. Specifically:

- **In** the **high-end range, BMW leads in immediate sales,** with **SUVs/Crossovers and Sedans** as the **preferred body types.**
- **In the upper-mid range**, **Ford stands out**, **with strong performance** from **SUVs/Crossovers, Pickups, and Sedans.**
- For the **mid-range**, **Chevrolet** vehicles top the list, **particularly in SUV/Crossover, Pickup, and Sedan** segments.
- In the **affordable segment**, **Honda** vehicles are the **most likely to sell instantly**, **with SUVs/Crossovers and Sedans** again **showing strong demand**.

Whether shoppers are prioritizing affordability, reliability, or versatility, SUVs and Sedans remain consistently desirable. For sellers, aligning inventory with these high-performing segments can lead to faster turnover and more strategic pricing decisions.

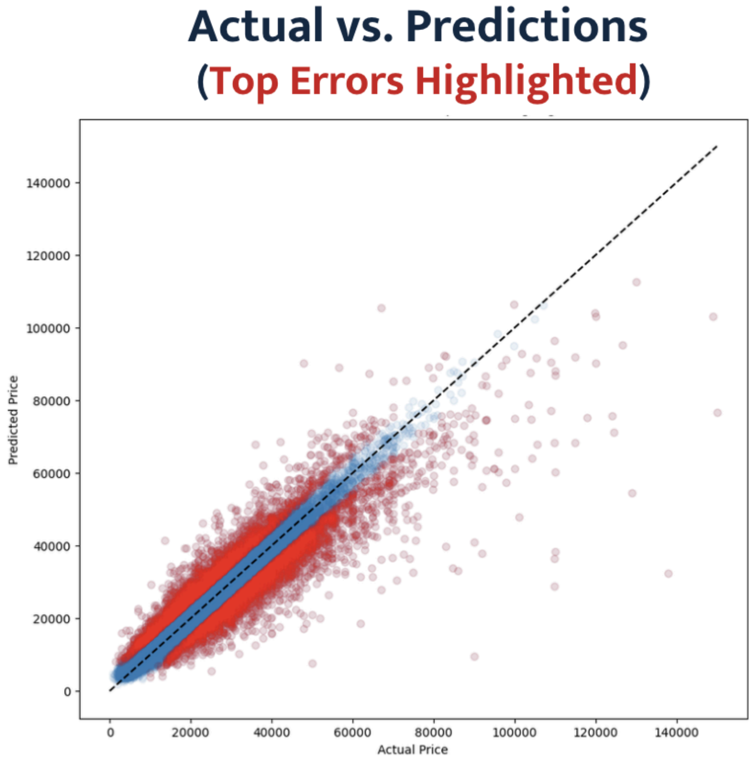
**Machine-Learning**

When addressing this project's machine-learning component, we found that **the Nearest Neighbor's Regression model best fits** our dataset, compromising both numerical and categorical values. **To accurately prepare our dataset** for modeling, **we encoded** our **categorical data into numeric** values. After **split**ting our **dataset into a 70:30 ratio for** our **train and test set**, we **then scaled** our dataset **to account for differences in scale**. **For example, the cars' age versus their respective mileage would be on completely different scales.**

After training and testing our machine-learning model, **we discovered that our model could predict car prices with a high accuracy rate**, explaining 90% of the variation in our data **($R^2$ = 0.90). When** the model was **predicting incorrectly**, **it was off by approximately $2301.07 on average**. In other words, the absolute mean error was 2301.07, and concurrently, this gave us an overall mean squared error of 13,523,612.81.

**To find the best parameters** for our machine-learning model, **we ran a GridSearchCV** to assist us with optimizing our model. In machine learning, **parameters are almost "settings" that** allow you to **control how the model** is processing data and **making its predictions.** When it comes to our Nearest Neighbors model, we are mainly looking at three parameters here: the number of neighbors, weight, and "P", which is just a distance metric. After running our GridSearch, **we found that the** "best" or **most effective parameters are as follows: number of neighbors = 5, P=2 (Euclidean), and weight = distance.**

Let’s break these parameters down. In **Nearest Neighbors, our model predicts car prices by analyzing features** (or variables) **of other cars** (or data points) **that are most similar to the car whose price we are trying to predict.** Following this logic, the **GridSearchCV found** that the **best number of “neighbors”** to look at a time **was five**. Furthermore, **the weight being “distance”** really **means** that **the model is putting more “weight”** or importance on data points that are closer in similarity versus other ones that are farther. As for “P,” we found that the model default P=2, the Euclidean distance metric, made the most sense here.



Actual vs. Predictions
(Top Errors Highlighted)

As pictured above, **our model performs** reasonably **well for the majority of cars** in our dataset. The **blue points,** which **represent 90%** of our data, **densely packed around the dashed line, indicate** that **most predictions are fairly close to** the **actual listing price.** The **red points,** on the other hand, **account for only 10%** of the data but appear more prominent because they are more spread out, **representing cars where the prediction error was much larger.** This highlights **our model’s** accuracy, with most **predictions being** tightly clustered and **reasonably accurate.** Importantly, **our model tends to underestimate actual prices more frequently** than it overestimates them. **This is reflected in the fact that many of our most significant errors fall below the dashed line.**

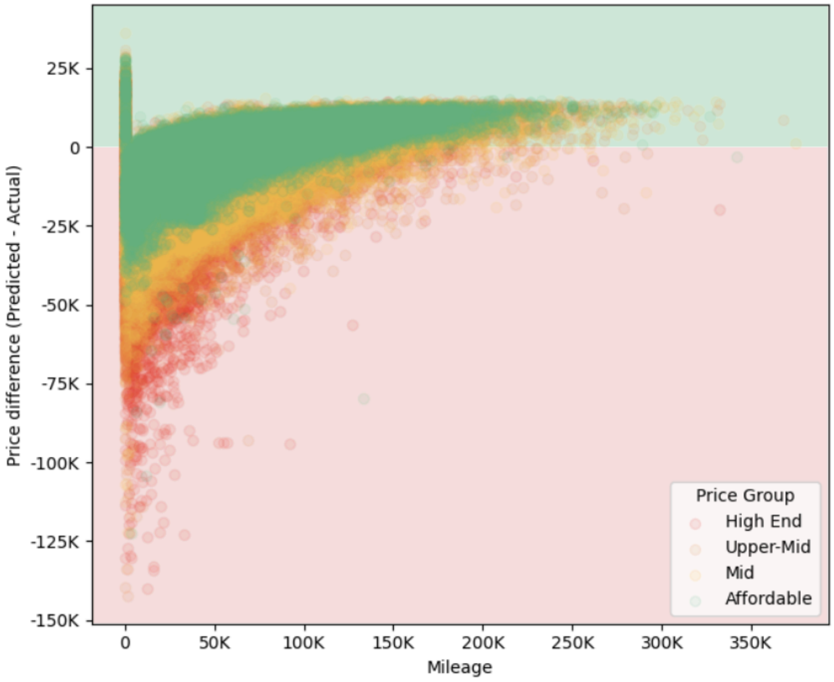
**Upon further analysis**, the **red data points mostly consist of high-end luxury cars**, **particularly Audi, BMW,** and **Mercedes**-Benz. Even **though these cars make up a smaller percentage** of our dataset, **they are overrepresented among the largest errors. One possible explanation is** that **these brands manufacture** a **wide range of vehicles from luxury** models with premium features **to** more **affordable**, family-oriented sedans and SUVs. **Convertibles and coupes** also **showed a greater price variance**, likely due to the variety of vehicle style, condition, and quality. **This variability may make it harder for our model to generate consistent predictions. On the other hand, more affordable brands** like Hyundai, Mazda, and Honda, tend to **show smaller errors,** which makes sense **given their larger representation** in our dataset **and more consistent market pricing.** These brands generally manufacture vehicles with fewer extremes in trim level, performance, or luxury add-ons. **Similarly, hatchbacks and wagons show less variation between** the **predicted and actual prices,** likely **because they are more** common, practical vehicles with **relatively consistent** conditions and features.

## Re-Applying Our Model

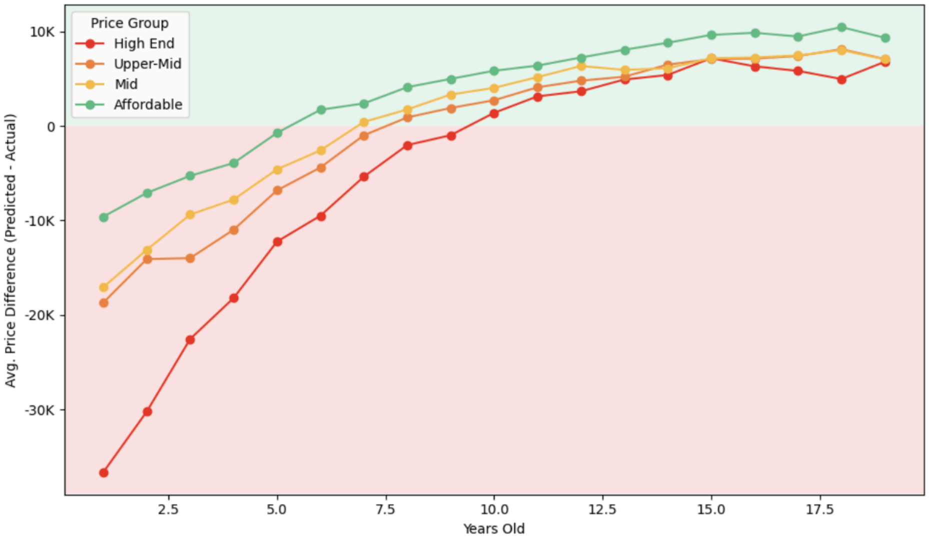
**Leveraging the predicted values** that our model gave us, **we wanted to find out if CarGurus** was **fairly pricing cars with an emphasis on** both **mileage** and **car age.** In hindsight, this means that **based on** CarGurus' inventory and **similar vehicles in the company system with identical features, we are putting** these car **prices to the ultimate test. Are customers potentially getting a bargain or maybe being overpriced? To understand this**, **we calculate** the price difference, we simply individually subtracted our actual car prices from our predicted **(Predicted Price - Actual Price). If** the difference came out to be **positive, then** the **customer would be getting a good deal.** Similarly, **if** the difference is **negative, then** the **car is possibly** being **overpriced.**

As shown below, **our results are grounded in intuitive reasoning. No matter** the **price group,** for **cars that have** a **higher mileage**, the more that they are sold at a **more cost-effective** price, and vice-versa. The **same logic holds true for older vehicles. The older** the car is, **the more that car is sold for a better price.**

## Scatter Plot by Price Group & Mileage



## Average Price Differences by Car Age



## Conclusion

**Our analysis revealed** that **SUVs, Crossovers, Pickup Trucks, and Sedans remain** the **most sought-after** body types across all price segments. Among the f**astest-selling brands, BMW, Ford, Chevrolet, and Honda consistently led** their respective tiers, **aligning** closely **with** these **high-demand** vehicle **categories**. One thing to note, **our machine learning model**—K-Nearest Neighbors Regression—**achieved strong predictive performance ($R^2$ = 0.90)**, reinforcing the reliability of these trends. **We** also **observed a clear association between higher mileage, older vehicles, and more affordable pricing.** These insights can offer valuable guidance for dealers, buyers, and sellers alike as they navigate the evolving dynamics of the used car market.

## GitHub: