Hayden Neer

Data Bootcamp

Professor Koehler

21 March 2024

**Predicting Sports Car Prices**

Setting the right sticker price can make or break a sports car retailer. A price that is too high gathers inventory, while a price that is too low erodes margins. For my final project, I developed a predictive model that recommends an optimal Manufacturer’s Suggested Retail Price (MSRP) for each sports car listing by learning from historical sales data. Using specifications—brand, horsepower, and 0-60 time—the model estimates what the market is willing to pay today. Dealers can then set MSRPs that fall within the sweet spot: high enough to protect profit, yet low enough to move cars off the lot quickly. Beyond day-to-day pricing, the model reveals which performance factors (and which badges) command the most significant premiums, helping purchasing managers decide which cars to stock. A reliable MSRP predictor, therefore, turns raw performance numbers into clear, money-saving guidance—aligning inventory, marketing, and sales strategy with real-world buyer behavior in the high-stakes sports car niche.

**Key Findings**

Exploratory work quickly revealed that three variables—manufacturer badge, horsepower, and 0-to-60 mph time—account for nearly all the predictive weight in this market. A brand captures the intangible prestige premium, horsepower quantifies raw performance, and 0-to-60 time serves as a concise proxy for overall engineering efficiency. When these three features alone feed a tuned Decision-Tree regressor, the model prices cars to within ≈ approximately $20,000 of their actual sale value, achieving the best performance of all algorithms tested and a practical accuracy band for setting an MSRP in day-to-day dealership operations.

**Dataset Description**

The analysis is based on the **“Sports Car Prices”** dataset, published on Kaggle by **Rungthip Kiattisak,** which provides real-world sales data. The file strikes a rare sweet spot: it is large enough to be statistically valid, encompassing over a thousand listings across **38 marques, more than 180 model trims, and production years spanning from the early 1970s through 2022,** yet compact enough to load and train multiple models within minutes. Each row couples hard-performance numbers (engine displacement, horsepower, torque, 0-to-60 mph time) with the actual sale price in U.S. dollars. The categorical brand and model labels are consistently spelled. Numeric columns are stored as strings but convert cleanly to floats. The blend of household-name badges (Porsche, Ferrari, Lamborghini) and ultra-low-volume exotica (Pagani, Koenigsegg) makes modeling difficult; therefore, manufacturers with fewer than 10 cars recorded in the dataset were removed.

**Exploratory Data Analysis**

Before any modeling can begin, the dataset must be examined, cleaned, and trimmed to ensure that the remaining observations accurately represent the segment a performance car retailer is concerned with. Listings priced above one million dollars—essentially auction-level hypercars—were removed because their prices behave more like collectibles than market-based inventory, which would unduly influence model coefficients. Electric-only manufacturers (Tesla, Polestar, Rimac) were also excluded, as battery technology and incentive structures drive their pricing dynamics in fundamentally different ways than those of gasoline vehicles.

A graph showing a distribution of sports car prices

AI-generated content may be incorrect.

Figure 1

The cleaned dataset has prices ranging from approximately $25,000 to $625,000, a **median of $114,000,** and a **mean of $174,000**. Figure 1 illustrates the interquartile range, which spans roughly $70,000 to $225,000.

A graph with blue dots

AI-generated content may be incorrect.

Figure 2

Figure 2 shows the relationship between the year a car was sold and the price a customer paid for the vehicle. Examining the scatterplot, there is no discernible relationship between the two variables. While one would expect retail prices for cars to increase over time, this is not immediately apparent in the plot. Due to the lack of a discernible correlation between price and year, the latter will not be included in any models to predict price.

A graph of different colored columns

AI-generated content may be incorrect.

Figure 3

Figure 3 illustrates the clear relationship between the manufacturer (make) and the price of a sports car. Looking at the chart, each brand has a distinct price range, with brands like Ferrari and Lamborghini at the top of the market, while Nissan and Chevrolet represent the lower end of the sports car market. Based on the noticeable price difference between brands, the brand will be incorporated into models that predict prices.

A graph of a graph

AI-generated content may be incorrect.

Figure 4

Figure 4 illustrates an interesting negative correlation between price and the time it takes a sports car to accelerate from 0 to 60 miles per hour. Vehicles that can perform this acceleration in less time (accelerate faster) are the most expensive cars. This is as to be expected, as a substantial reason people purchase sports cars is to race and experience the acceleration while driving. Based on the clear relationship here, while not linear, the zero-to-sixty time of vehicles will be included in models created.

A graph of a number of blue dots

AI-generated content may be incorrect.

Figure 5

Similarly, Figure 5 illustrates a clear positive relationship between a sports car’s horsepower and the price customers are willing to pay for it. This is also to be expected, as horsepower is frequently leveraged as a marketing tool; brands like Bugatti boast horsepower values exceeding 1,000, which is far above the market average. Given this evident relationship, horsepower will be included in models designed to predict the prices of sports cars.

A graph with numbers and points

AI-generated content may be incorrect.

Figure 6

Alternatively, Figure 6 shows a little relationship between the number of cylinders within an engine and the price of the car equipped with that engine. This is unexpected, as adding engine cylinders is generally increasingly expensive. The lack of clear correlation here may be the result of low-end sports cars having cheap high-cylinder engines, while high-end cars often contain extremely high-quality and efficient, low-cylinder engines. Due to the poor correlation, engine size will not be considered in models used to predict price.

A graph of torque versus price

AI-generated content may be incorrect.

Figure 7

Much like engine size, Figure 7 shows a weak positive correlation between the pound-feet of torque a sports car can produce and its price. This might be because the amount of torque a vehicle needs to navigate at high speeds effectively has an upper limit, and high-end cars do not surpass low-end vehicles in this regard. Torque will not be considered when creating models to predict price because of this.

**Models and Methods**

With the data cleaned and key features selected, I trained a series of increasingly flexible models to determine how close each could approximate real-world prices. I progressed from the most transparent (linear regression) to a local method (KNN), and then to rule-based and decision tree models, benchmarking each step against a simple “average price” baseline. The goal was not only to achieve a low error rate, but also to gain a clear understanding of which approach strikes a balance between accuracy, interpretability, and resistance to overfitting.

**Linear Regression Model**

I started with **linear regression** because it’s the simplest and most transparent way to convert specifications into dollars. After fitting the model, I checked the error on both the training set and the unseen test set. The training root mean square error (RMSE) came out to **≈ $60,000**, and the test RMSE landed at **≈ $56,500**. That tight gap indicates that the model is learning the actual signal, not just memorizing the training data. For context, my baseline—always guessing the average car price—had an RMSE of **$143,000**, so even this simple linear model cuts the typical pricing error by roughly **60 percent.** Put differently, brand, horsepower, and 0 to 60 time alone let me price a car to within about fifty-odd thousand dollars.

Permutation-importance scores highlight why the model performs well: the manufacturer badge (0.68) and horsepower (0.65) are by far the two strongest predictors. At the same time, 0‑to‑60 time (0.03) adds only a marginal boost once raw power and brand reputation are known. This ranking aligns with market intuition—buyers pay first for the name on the hood, then for the horsepower underneath, and only slightly for the stopwatch figure. The linear model’s coefficients echo the same story, confirming that most of its $56 K prediction error comes from variations that those three variables can’t fully capture, such as limited‑edition trims or condition details not present in the dataset.

**K-Nearest Neighbors Regression Model**

The K-Nearest Neighbors regressor (with the optimal k value determined through cross-validation) takes a distinctly different approach from the linear model. Instead of fitting a global equation, it prices each car as the average of its closest peers in the feature space. That “let the data speak” approach paid off. On the training set, the model achieved an RMSE of roughly $21,000, and on the unseen test set, it held steady at $23,800—less than one-fifth of the $143,000 baseline error and less than half of the linear model’s $56,000. The tight train-to-test gap indicates that the algorithm isn’t simply memorizing specific listings; it is genuinely capturing the local price structure in the neighborhood defined by make, horsepower, and 0-to-60 time.

Permutation importance confirms why KNN works so well in this context. If I randomly shuffle horsepower, the test error jumps the most (score ≈ 0.83), followed closely by make (0.78); together, they anchor the similarity search. The 0-to-60 metric contributes about half as much (0.44), consistent with its partial redundancy once horsepower is known; yet, it remains helpful for distinguishing between two cars with similar power but different acceleration profiles. Because KNN weights every feature equally after scaling, it naturally leverages these high-signal variables to pull each query car toward appropriately priced neighbors, delivering highly accurate MSRPs without a complex model structure.

**Decision Tree Regression Model**

A graph with blue and orange dots

AI-generated content may be incorrect.

Figure 8

The decision-tree regressor allows me to trade the straight-line assumptions of linear models for a set of straightforward “if-then” price rules, without the variance blow-up that can plague deep trees. I used a grid search over depths 1–19 and plotted train-versus-test-set MSE (Figure 8); both curves dropped sharply until about depth 9, then flattened, with the test error creeping upward past depth 12—classic over‑fitting. Locking the tree at max\_depth = 9 struck the right balance: simple enough to generalize, complex enough to capture brand‑by‑power interactions. On unseen data, the model achieved an RMSE of roughly $19,300, trimming another 20 % off the already‑strong KNN error and beating the naive average‑price baseline by more than 7 times. The tight train‑to‑test gap confirms that the depth‑tuning step successfully prevented the tree from memorizing individual listings.

Feature‑importance scores extracted from the tuned tree echo earlier findings but add nuance. The first split is almost always based on make, instantly separating ultra-premium badges from mainstream ones; subsequent levels pivot on horsepower, and only the deepest nodes reference 0-to-60 time, usually as a tiebreaker between two cars of similar power. Because a decision tree is inherently interpretable, dealers can read these splits as actionable pricing guidelines—for example, “if make = Ferrari and horsepower > 650 HP, start near $400 K regardless of year.” Compared to the black-box feel of KNN, the tree offers nearly the same predictive accuracy while also serving as a transparent pricing playbook.

**Random Forest Regression Model**

Although the single decision tree already priced cars with impressive accuracy, I still trained a Random Forest regressor—50 trees, each capped at a depth of 20—because ensembling is intended to smooth out the high variance that individual trees can exhibit when they analyze only one version of the data. By averaging many decorrelated trees, a forest typically trades a small amount of bias for a significant reduction in variance, so in principle, it should either match or exceed the performance of the best single tree while being less sensitive to minor data quirks.

In practice, the forest landed at approximately $13.3 K RMSE on the training set and about $20.0 K on the test set. This is essentially on par with—though slightly worse than—the tuned depth-9 decision tree’s $19.3 K test error. The nearly identical train-test gap confirms that the forest is generalizing well; the slight underperformance likely stems from my deliberately conservative hyperparameters (only 50 estimators and a depth ceiling of 20). A broader grid search over n\_estimators, max\_depth, and min\_samples\_leaf could still uncover a configuration that surpasses the single tree. Still, even with these settings, the ensemble demonstrates that the results are stable: aggregating multiple trees does not degrade accuracy. It provides a more robust, lower-variance fallback option if future inventory differs from the current dataset.

**Conclusion**

Starting from a $143 K baseline error, each modeling layer sharpened the price lens: linear regression cut the miss to $56 K, K Nearest Neighbors dropped it to $24 K, and a tuned decision tree settled at approximately $19 K RMSE—accurate enough to set an MSRP within 10% for most vehicles on today’s lot. Ensemble trees (Random Forest) confirmed that the signal was robust, matching the single tree error while offering additional protection against future data drift. These results demonstrate that even a modest dataset of 840 listings can yield a practical, data-driven pricing tool.

That said, the model remains blind to factors it has never encountered. Mileage, service history, option packages, and regional demand are obvious omissions—each can significantly impact the price by tens of thousands. Collecting a larger, multi-source dataset that includes those variables and at least double the number of cars would provide the algorithms with much richer context and likely push the error well below the $15,000 mark. Finally, retraining quarterly will enable the system to learn real-time shifts in the performance car market, transforming a strong proof of concept into a living asset for any dealership’s pricing strategy.

**Works Cited**

Kiattisak, Rungthip. “Sports Car Prices Dataset.” **Kaggle**, 5 Sept. 2021, [www.kaggle.com/datasets/rkiattisak/sports-car-prices-dataset](http://www.kaggle.com/datasets/rkiattisak/sports-car-prices-dataset). Accessed 12 May 2025.