Grayson Nickel
9/28/2025

The Best Value Cars of 2025

Project 2: Classification

The problem

For this project, I am asking a very similar question to project 1 using the same data set. My question is what are the best value cars of 2025. Last time, I asked essentially the same thing but did not do a good job at calculating the best valued cars. All I did was divide total price by horsepower and determined that the least amount of money per horsepower was the best deal. That way is very flawed so this time I will be using a decision tree considering vehicle price with a limit to how high it can be so we don't see very expensive vehicles like the Cybertruck again. The decision tree will also consider horsepower with a minimum amount required so the list isn't full of foreign, unsafe and not street legal Tuk-tuks and the like. Essentially this project's problem is a remaster of my project 1's question using a decision tree to classify each car on the list as either good value or bad.

The Data

Same as project 1, this data set is for 2025 including over 1,200 cars. The dataset is called "Cars Datasets (2025)" posted by Abdul Malik on kaggle. This dataset is for free use as long as it's not malicious or for profit. This data set has 11 columns in total covering the following data:

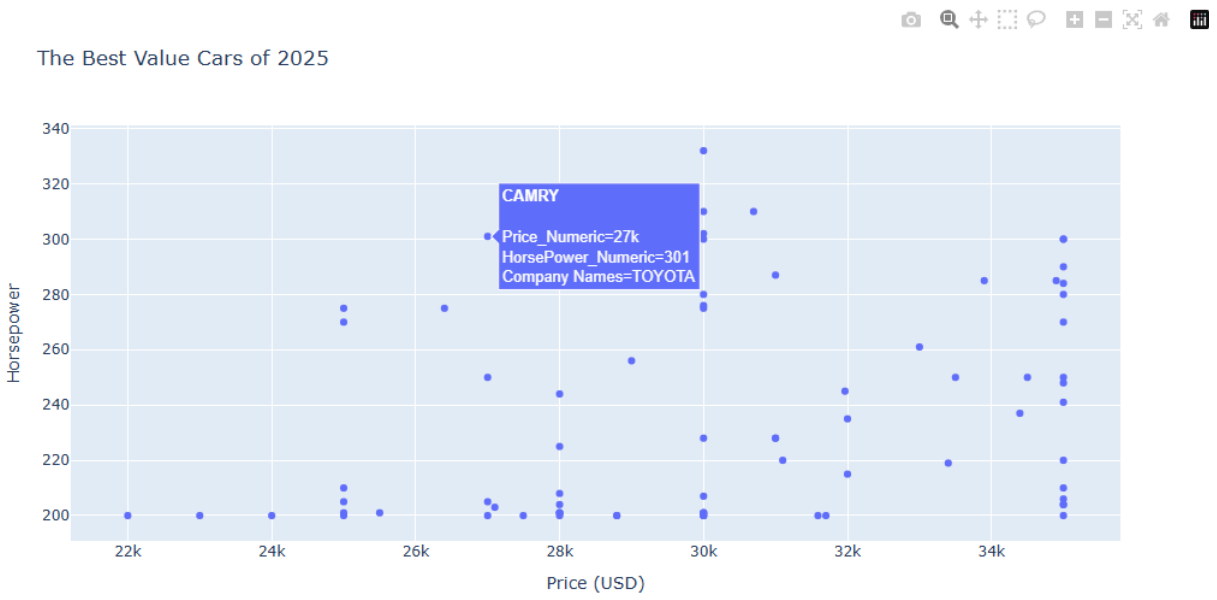
Car Company Names: The manufacturer or brand of the car.
Car Models: The specific name or series of the car.
Engine Types: Information on engine specifications .
CC/Battery Capacity: Engine displacement in cubic centimeters or battery capacity for electric cars.
Horsepower (HP): The power output of the car's engine or motor.
Top Speed: The maximum speed the car can achieve.
0-100 km/h Performance: The time it takes for the car to accelerate from 0 to 100 km/h.
Price (in USD): The car's price listed in United States dollars.
Fuel Type: Specifies whether the car uses petrol, diesel, electricity, or hybrid fuel systems.
Seating Capacity: The number of passengers the car can accommodate.
Torque: The rotational force the engine generates.

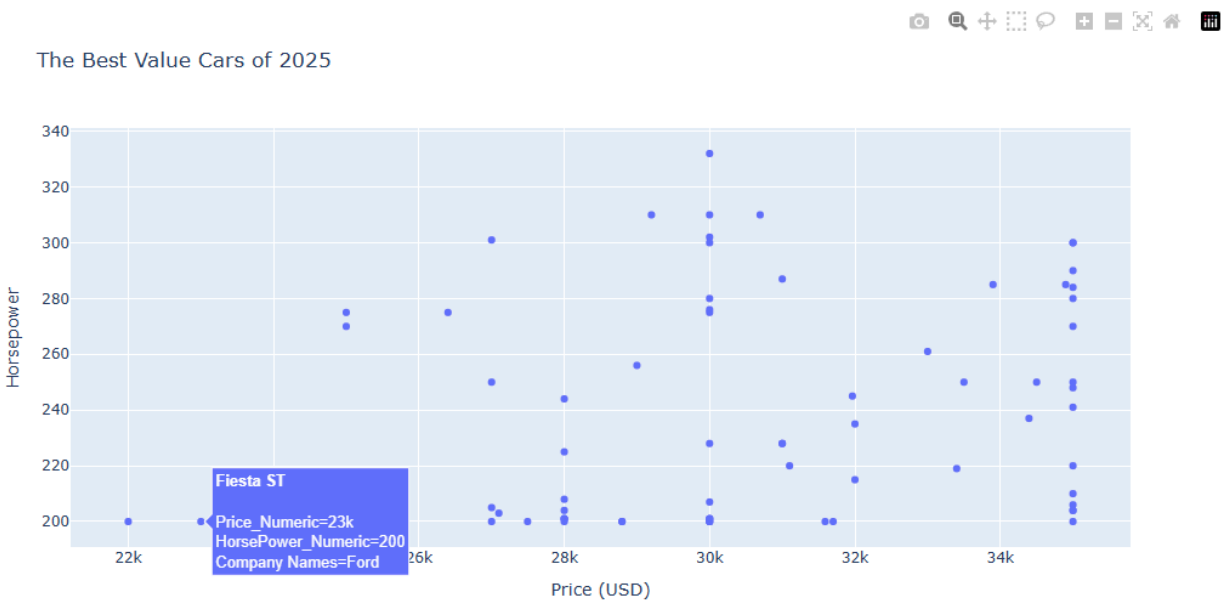
Pre-processing

My first step was to set up my environment with all the necessary imports and directories for the data set. Next, I cleaned the data for the price and horsepower columns as they are stored in the data set as strings with money signs, commas, and text which would get in the way of calculations. Next, I removed any rows with missing data. Next I set up my decision tree to make sure everything works, which it did. Next, I calculated which column value I should place as the root node of the decision tree. I did this by calculating the impurity of 'Price_Numeric' and 'HorsePower_Numeric' using Gini impurity. With that, I found that 'Price_Numeric' has a lower Gini impurity and is the best fit for the root node. Finally I created a visualization of ALL of the good value candidates using Plotly which allows you to hover over each dot and see more information about each specific car without overcrowding the scatterplot.

Data Understanding/Visualization

While exploring my improved list of good value cars, I immediately noticed a much cleaner set of data with more realistically attainable cars with my power and price limits in place. I also have a new appreciation for Plotly and how powerful of a tool it can be when trying to plot large data while still including details about each specific point on a plot. My Plotly visualization is a scatterplot containing every single car classified as good value, while including make, model, price, and horsepower for each one of those cars and it doesn't look the least bit crowded. You simply hover over each point to see more information about it.





Modeling

The classification model I used for this project is a decision tree containing a check if the price is less than $35,000. False is a leaf setting the item to not good value while true branches to a check seeing if the car makes at least 200 horsepower. This leads to two leaves, true will set the item to good value and false sets it to false. I determined that ‘Price_Numeric’ should be the root node of the decision tree by individually calculating the gini impurity of ‘Price_numeric’ and ‘Horsepower_Numeric’ and then manually comparing to see which impurity was lower. The price was lower impurity so I used that first for efficiency's sake.

Evaluation

Overall, I say my model performs very well. I calculated the gini impurity once so it’s not like I ever need to run those lines again and the decision tree uses the lower impurity column first and only has two decisions to make total to determine if a car is good value or not. So that paired with this data set only having around 1,200 cars total makes waiting times on execution non-existent.

Storytelling

I learned what cars from 2025 are a good value considering my personal preferences of a maximum price of $35,000 and a minimum of 200 horsepower. From that list, I particularly like the Toyota 86, GT86, Nissan 370Z, and the Ford Fiesta ST. That information actually checks out because those are all great value cars that have been on my radar for years considering sporty performance and price. With that, I was able to answer my question of what the good value cars of 2025 are. I was even introduced to some new cars I never considered like the Hyundai Veloster.

Impact Section

With my more accurate list of the good value cars of 2025, I think this list will have a positive impact on people in the market for a new car. Of course the decision tree was made with my preference for maximum price and minimum horsepower in mind but this list should still provide a good baseline for everyone else. I can't really think of any genuine negative impacts this list could have because all of the more foreign and dangerous or outrageously expensive cars are no longer on the list which makes all of the options more viable for other people looking for a brand new car. However one negative thing to consider about this list is that it cannot predict the reliability of these cars as brand new cars have yet to prove themselves over the test of time and use. That is a rather important thing because most new cars these days use more and more plastic and more and more complex designs which could reduce durability and reliability. And if you're already on a budget concerning a new car, then you may not be able to also afford expensive repairs if the car ends up being unreliable.

References

The dataset: https://www.kaggle.com/datasets/abdulmalik1518/cars-datasets-2025
Seaborn API: https://seaborn.pydata.org/api.html
Plotly Library: https://plotly.com/python/
Video tutorial: https://www.youtube.com/watch?v=LDRbO9a6XPU

The Code

All of the project files including the code and csv can be found here:
https://github.com/GMNICKEL/Projects/tree/main/Python%20classification%20project%202