Predicting Used Car Prices

(COMP3125 Individual Project)

Bjordi Ismailati
Data Science Fundamentals

***Abstract*—This project analyzes crucial factors that influence the price of used cars by using real market data. The focus for this project has been on mileage, brand, age, and specific features of each model. For finding these extra features and details on the cars, the project used decoding VIN Python code. This enhanced the usability of the code to help with visual representation and give more data for training the model that goes on to predict the future price of the used car brands.**

Keywords—used car prices, price prediction, machine learning, regression.

I. INTRODUCTION

Many people, including students, myself, and others, will need to buy or sell a used car at some point. We know that it's not always easy to know if the price is good or bad, as some cars lose value quickly, while others keep their value for longer. Just a while ago, when trying to sell my car, I had to face a question: "What is the value of my used car, and how will it be affected in the future?"

For this project, I analyzed my dataset and focused on key factors such as mileage, age, and brand to work out how the value of a car gets affected by these more general factors. Then I reviewed the determining factors that features like fuel type, drive type, body class, and displacement have on the value of a used car.

To find these relationships, I cleaned the dataset and used libraries through code so I could gain more details for the cars in my dataset and have a more complete dataset. I created visualizations to display how car value is affected by mileage or by the car's age. Other visualizations also include how different brands have better value management than others and display which those are from my dataset. Most importantly, the goal of this project is to develop a machine learning model to estimate used car prices in the future.

II. DATASETS

Dataset:
https://www.kaggle.com/datasets/doaaalsenani/usa-cers-dataset.

The dataset used for this project has scraped the data from the well-known site of auctionexport.com, and it is a record of a certain amount of their sales. These sales focus on multiple brands, and it only contains cars sold in North America. The dataset only had a record of the car's brand, mileage, age, state, and VIN number. As it was missing specific car features, I ended up modifying the dataset by implementing code that had access to the well-known NHTSA site, and through VIN decoding, I was able to include specific features of the cars in my dataset, making for a complete dataset.

A. Character of the datasets

This cleaned dataset contains:

- price: the price the car was listed/sold for.
- brand: car manufacturer.
- model: model name.
- year: manufacturing year.
- mileage: total miles driven.
- Engine Displacement (L): size of the engine.
- Cylinders: number of cylinders in the engine.
- Drive Type: like FWD, RWD, AWD, etc.
- fueltype: type of fuel the car uses
- Body Class: SUV, sedan, truck, etc.

This dataset contains almost 2500 different cars sold from the auctionexport site. It contained 23 different brands that had additional features included in their dataset. As mentioned above, I needed to make changes to the dataset and add additional features. During the data cleaning process, I needed to do some normalizations so that when the data that was gathered from decoding the VIN number arrived, it would arrive in the format that would make it easy for me to understand and use. This included normalizations such as when looking at the drive type in the NHTSA site, the code would only deliver the car's feature in its shortened form, so FWD for front wheel drive or AWD for all wheel drive, and other normalizations like that. Other normalization included features like body class, where the site would return multiple responses for the feature, so I made it normalize to a certain shorter version to make it easier to read the dataset.

I. Methodology

Before moving on to viewing the visual results, first, I used the VIN decoder site NHTSA and the request library to access the site and deliver features to the dataset.

Next, I removed entries with values registered as "Unknown" or those with missing values. I also engineered new features to enhance model accuracy. Additionally, I grouped vehicles by age into a yearrange feature for use in visualizations. This was done in the format 2001-2005 and as follows for years after or before. These steps ensured a cleaner, more structured dataset suitable for modeling and visualizations.

Moving on, I created several visualizations using the cleaned dataset:

- Price vs. Mileage: I used a plot with a regression line to show how a car's price tends to decrease as

mileage increases. This answers the question of how mileage impacts value.

- Brand Value Over Time: I plotted the price trend by age for the top five most represented car brands in the dataset. This visual shows me which brands hold their value better over time.
- Price vs. Age: Another plot like that of mileage vs price but focusing on age, which showed the regression line for the car's age. This helps examine the relationship between age and price in general.

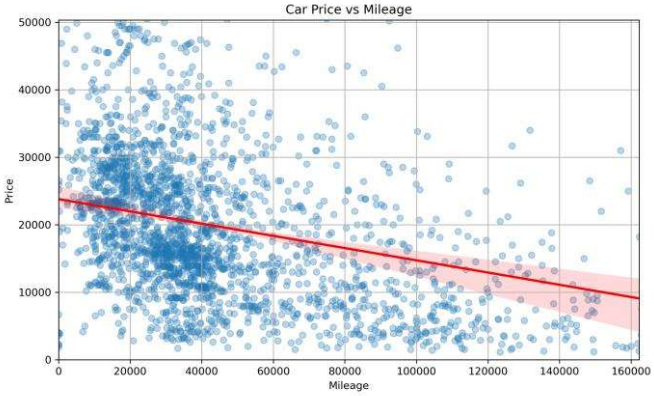
These plots not only made trends easier to understand but also supported the insights later used in the machine learning model.

To explore whether car prices can be predicted, I trained a machine learning model using a Random Forest Regressor. I chose this model because it handles both numerical and feature data well and provides solid performance (performance-the one thing I hated about this project, it would take so long to load all the data just to find I had something that needed a better way to display it). Prior to training the model, I processed the data by scaling the numerical features to standardize their ranges and applying dummy variable encoding to feature variables, such as fuel type, brand, and drive type, to make the data compatible with the algorithm. After training, the model also displays the mean absolute error (MAE), closer to zero the better, and r^2, which indicates how good the model is at determining the value for the certain dataset I loaded to it; the closer the value is to 1, the better the model predicted the price. These results show that the model can reasonably predict car prices based on features, though there's still some error margin.

III. Results

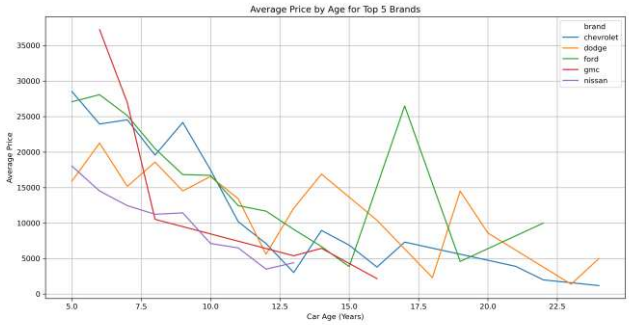
A. Price vs. Mileage

Our analysis shows a clear and steady downward trend between mileage and car price. As mileage increases, car prices decline. This result aligns with expectations and highlights mileage as a key factor in depreciation.



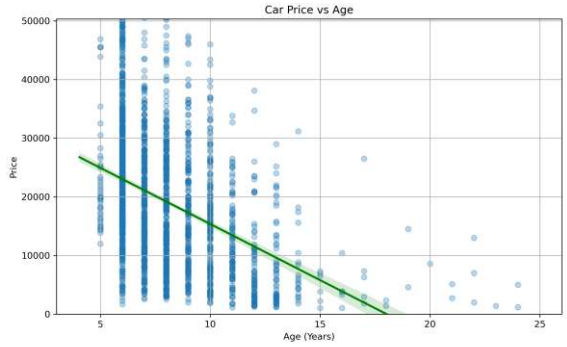
B. Value management by Brand

Line plots of average price by age reveal that all five top recurring brands from the dataset (Chevrolet, Dodge, Ford, GMC, and Nissan) generally experience a decrease in value as the vehicle ages. However, one interesting outlier appeared in Ford vehicles around age 17, where prices spiked above competitors. This could be due to a rare or mispriced data point messing up the average and highlights the importance of taking off outliers during analysis.



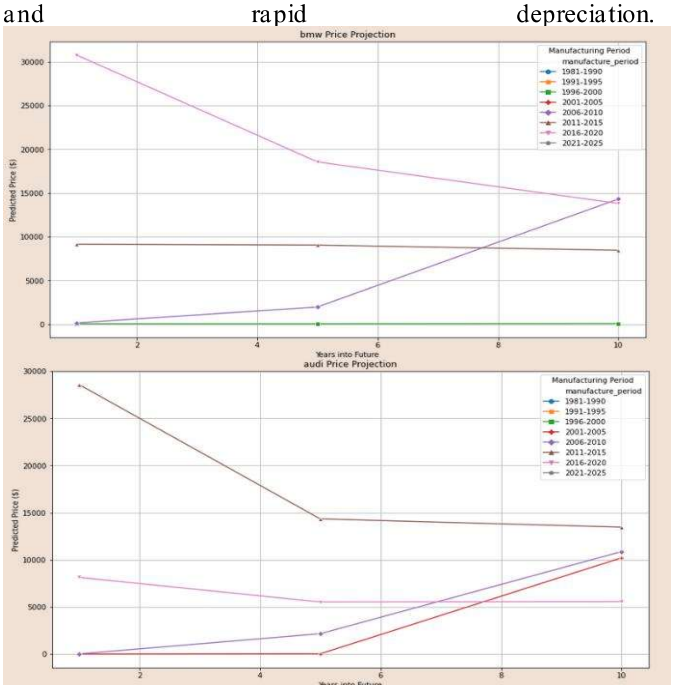
C. Price vs. Car Age

Regression plots of used car prices against age show a strong and consistent decline. The older the car, the lower its average resale price. This pattern reinforces that age is a major determinant part of depreciation.

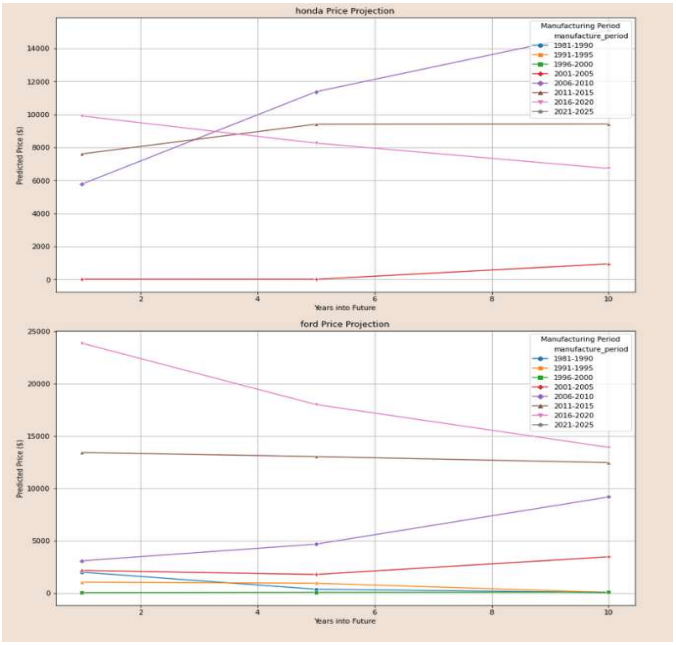


D. Model-Based Brand Insights (Question 4)

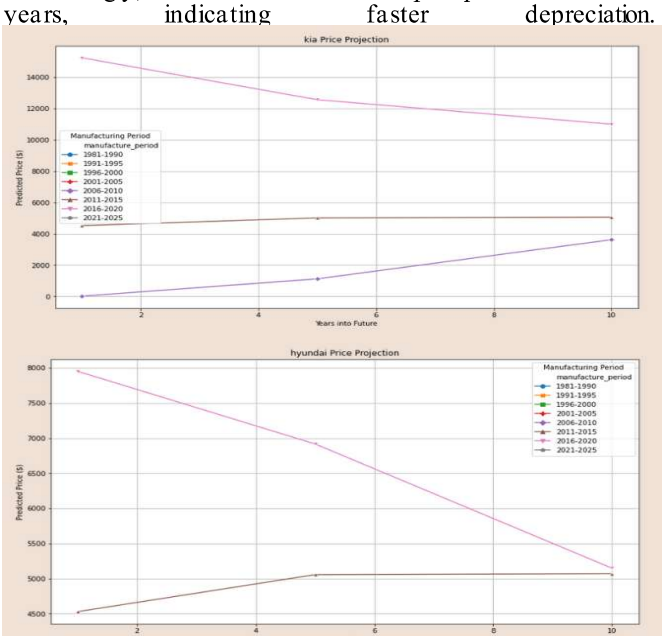
Using the trained model, the code predicted future car prices across 1, 5, and 10 years in the future. The results showed that brands like Lincoln, BMW, and Audi lose value very quickly in the first few years, likely due to luxury pricing and rapid depreciation.



On the other hand, brands like Acura, Ford, Mazda, and Honda retained their value better—especially among older vehicles.



Interestingly, Kia vehicles showed a steep drop in value across all years, indicating faster depreciation.



Also, the data displayed a decent performance from the model with an MAE of $3300 and an r^2 of 0.75. While the MAE is pretty big, that could be improved with more trends and more data being used by the model. The 0.75 shows a good strength from the model prediction.

```
Model Performance:
MAE: $3,353.55
R²: 0.75
```

IV. DISCUSSION

The results reveal depreciation patterns across multiple car brands, offering insights into used car valuation dynamics. Luxury vehicles like Lincoln, BMW, Lexus, or Audi experience rapid initial depreciation, losing significant value within the first 3-5 years for their newer brands. This aligns with industry observations that luxury cars often depreciate faster due to the high cost of purchase and maintenance

expenses. However, their depreciation rates generally stabilize after a 5-year period, suggesting they can be more appealing to budget buyers seeking premium features.

Other brands to note include Kia, which demonstrates a steep decline in value over time, reflecting perceptions of lower long-term reliability. In contrast, brands like Acura, Ford, Mazda, and Honda show great value retention, with older models maintaining or even appreciating in price. This trend may come from strong reputations for durability and affordability of replacement parts.

Future works

- The absence of famous brands in my dataset, like Toyota and Volkswagen in the results, indicates dataset limitations. Future work should prioritize broader data collection to ensure complete brand representation.
- The model assumes fixed annual mileage of 12,000 miles per year and linear depreciation, ignoring driving patterns or classic car appreciation. Integrating usage patterns and non-linear aging could improve accuracy.
- Economic trends like fuel prices going up or down, or EV adoption in the community increasing, and other similar trends were not accounted for. Including these variables would enhance the predictive strength of the model and give it more realistic flavors.
- Rare brands in my dataset produced unstable predictions due to limited data. This also could be avoided by a more complete dataset or by a better model that is better trained to find price without a lot of data.

V. Conclusion

Even though we already went over trends that the model could explore to be more secure in its prediction, it still highlights critical trends in used car depreciation, offering valuable data for buyers or sellers:

- Luxury brands experience rapid value loss within 5 years, making them high-risk purchases for short-term owners but viable options for long-term buyers.
- Non-luxury brands like Acura, Ford, Mazda, and Honda demonstrate strong value retention, with older models often maintaining or increasing in price. These brands are safer investments for cost-conscious consumers.
- Brands like Kia that experience steep depreciation warrant caution, suggesting buyers prioritize alternatives with better resale trajectories.
- Owners who want to stay away from wild resale values end up staying with brands like Honda, which have a good reputation for their resale value.

This project provides a great general understanding of used car values. While the model isn't perfect, it gives regular people a chance to view what they are getting themselves into in a market that can be tough to predict for depreciating assets like cars.

References

[1] "US Cars Dataset", www.kaggle.com. https://www.kaggle.com/datasets/doaaalsenani/usa-cers-dataset.

[2] "Car Export & Used Auto Auction, Buy Cars Online from USA/Canada Dealer," Auctionexport.com, 2025. https://www.auctionexport.com/?gad_source=1&gclid=CjwKCAjwwe2_BhBEEiwAM1I7sQezDxtNTE4mep-pIBoqZS9WioCEj9AdP8daev15-AbbzVOSEkemyRoCSfwQAvD_BwE (accessed Apr. 13, 2025).

[3] [1]"VIN Decoder | NHTSA," www.nhtsa.gov. https://www.nhtsa.gov/vin-decoder