AAP - Which car features contribute to sales?

Caden Goodwin

MGSC 310

Dec 15, 2023

Intro

Since the beginning of commercial car production, car manufacturers have had the question how they can get more sales. Aside from marketing and brand appeal, what distinct features in a car lead to higher sales? By taking into account these factors car manufactures can boost sales for their company. This question will be answered in my analysis backed by plots and models sourced from a car sales dataset from kaggle.

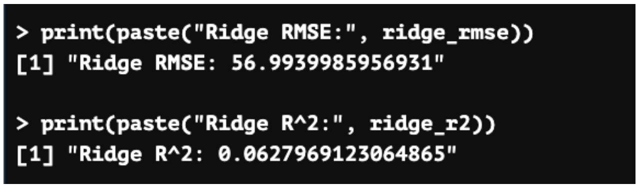
In my dataset I will be looking at the variables "Sales_in_thousands (the target variable), year_resale_value, Price_in_thousands, Engine_size, Horsepower, Wheelbase, Width, Length, Curb_weight, Fuel_capacity, Fuel_efficiency, and Passenger count", along with the a variable I created, "variable Price_to_Horsepower_Ratio" to make some insights about the importance of features and their overall influence on car's sales. Manufacturers can use this data to prioritize their engineering team on these influential features. For example if more fuel efficient cars tend to sell more, manufacturers can use this information to push towards making EVs.

Pre Processing

I first had to preprocess my data to make sure my analysis was reliable. I converted my categorical variable "Vehicle_type" to a dummy variable so it would work with my model. I also omitted the missing values.

Model

I ran into challenges when creating a model. I tried a linear regression model which was getting a RMSE around 80 with R^2 nearing 0. This was not good because even though some cars made hundreds of thousands of sales, this means it was still off by around 80 thousand on average. I tried a GBM model and was getting a similar poor result. I tried adjusting the tree number and depth and was able to get a slightly lower RMSE around 60, but that still wasn't ideal and modifying values in a GBM usually leads to overfitting. After a lot of trial and error, the best model I got was a Ridge regression model using cross validation with a RMSE of about 56.9 and a R^2 of about .06. These results still aren't great, but after trying just about every model I possibly could, I was able to see a decent improvement with ridge and cross validation. I included the other 2 most accurate models in my code which were a cross validated lasso model and a Random Forest model, which were slightly worse than my ridge results but comparable. I don't think it's possible to have an ideal outcome for an accurate model with my dataset because there's so many variables that don't contribute to sales in a linear way.

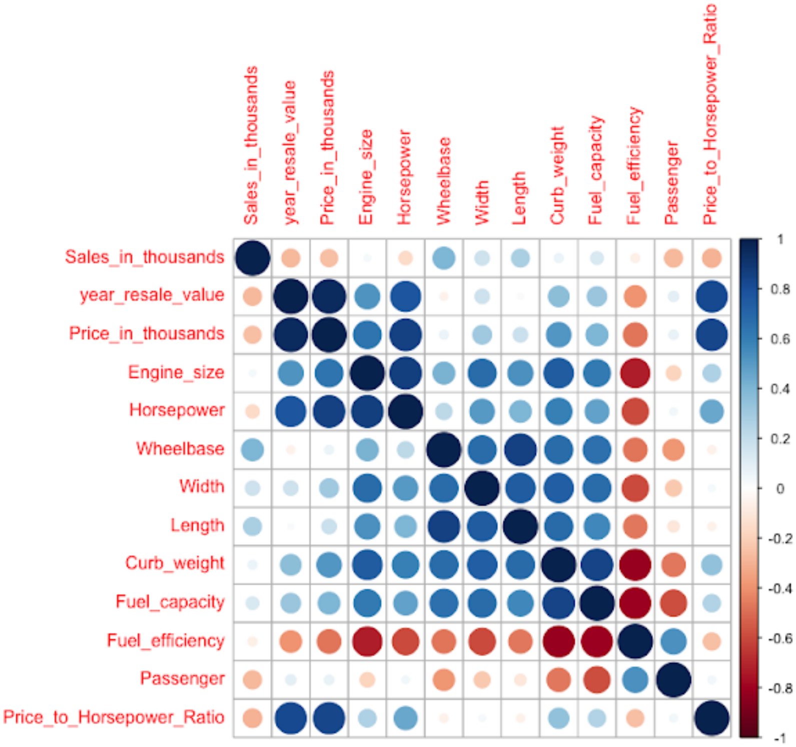


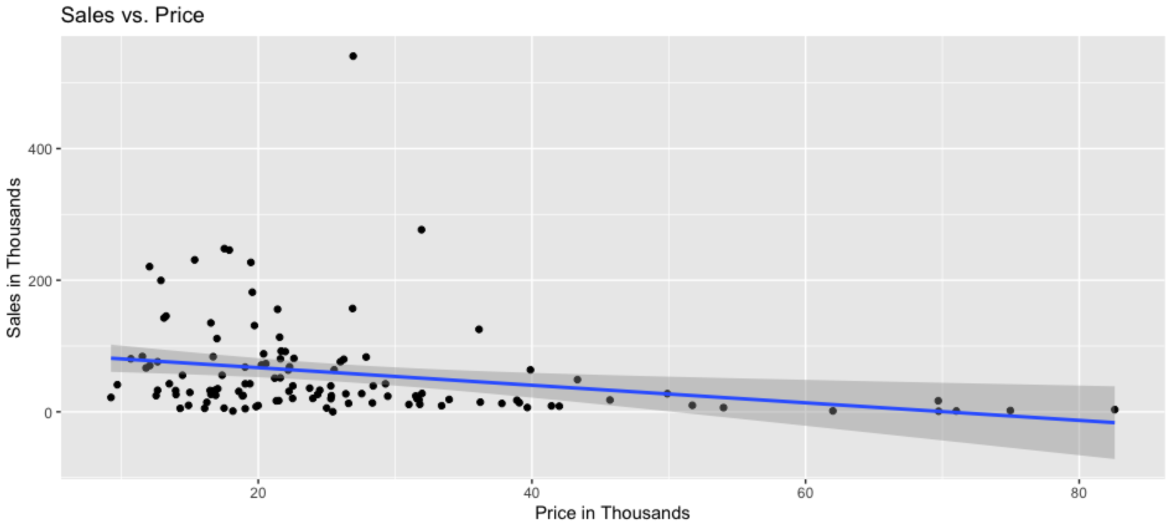
```
> print(paste("Ridge RMSE:", ridge_rmse))
[1] "Ridge RMSE: 56.9939985956931"

> print(paste("Ridge R^2:", ridge_r2))
[1] "Ridge R^2: 0.0627969123064865"
```

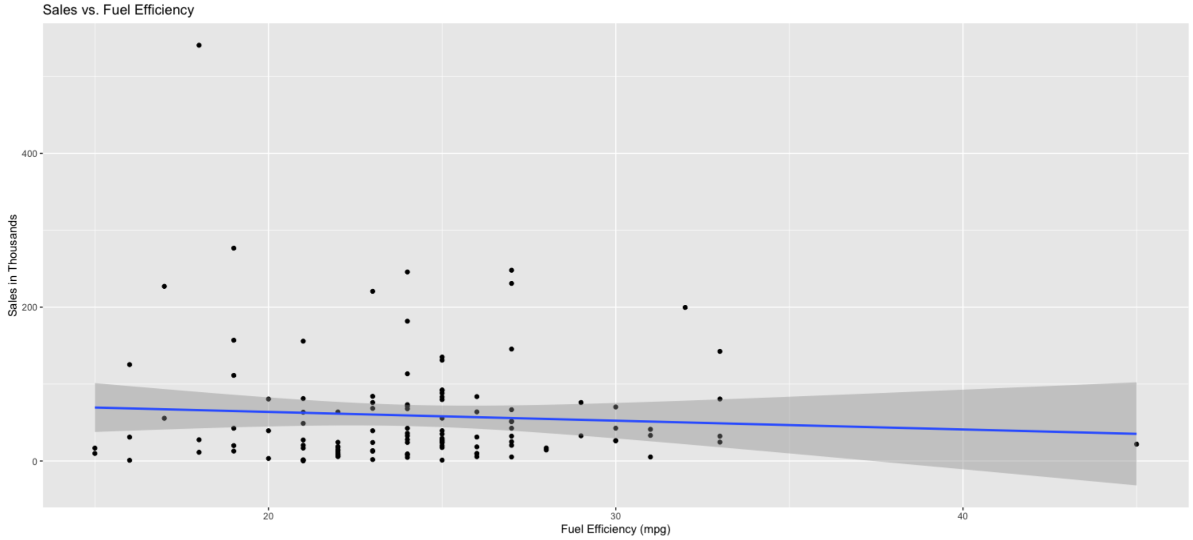
Plots + Analysis

The first plot I created is a correlation matrix which not only shows us how strongly correlated car sales are to variables but also how correlated other variables are to each other. Some are pretty obvious like how engine size is very strongly correlated to the car's horsepower, but our question is to see what influences sales. One of the insights I got from this correlation matrix is how wheelbase is a factor in higher sales. Usually higher wheelbase cars are trucks, so this can be telling us that trucks tend to do better in sales. Somewhat surprisingly, horsepower and car sales had a slight negative correlation meaning there might not be that much of a demand for high horsepower cars from the average consumer. Another thing that was somewhat surprising is fuel efficiency had a very slightly negative correlation to sales. This dataset is somewhat old, so maybe at that time there was not that much of a demand for fuel efficient cars. Overall this plot is very important and shows us the relationship between the car features variables.

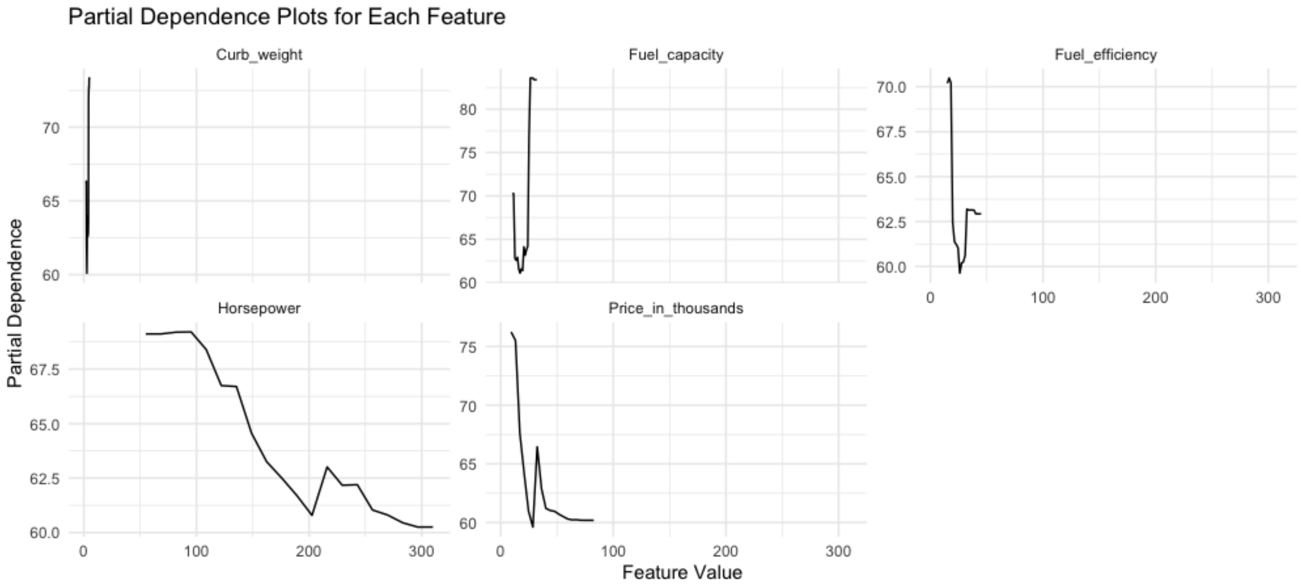




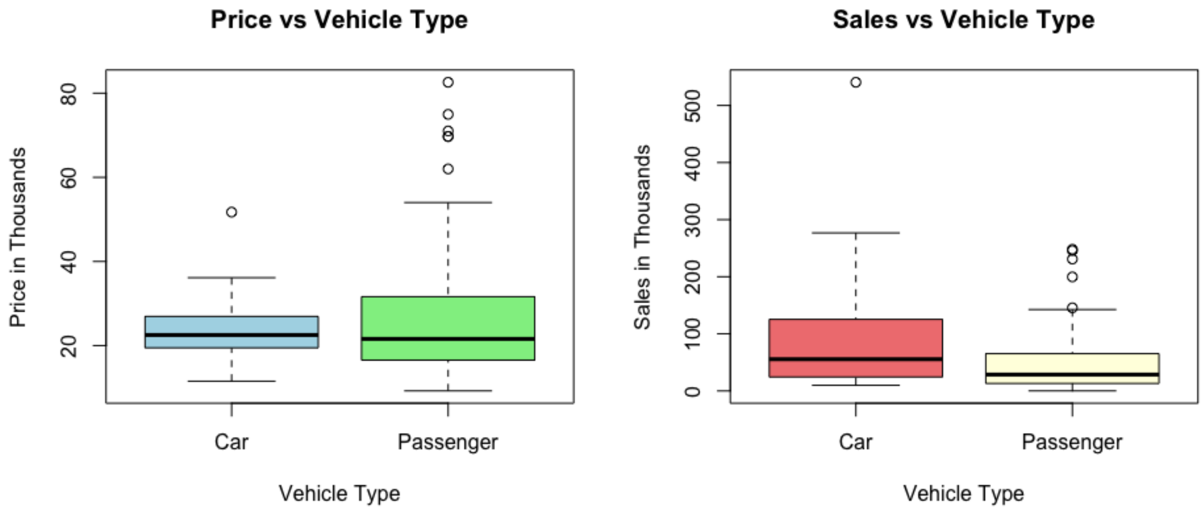
This scatter plot shows us that there's a slightly negative correlation between car price and sales in thousands of cars. I included a trend line to easily see the downward trend of sales at higher prices. The most obvious takeaway from this plot is when the price of a car increases then the sales generally decrease. Most of the cars in this dataset were pretty mainstream, average consumer cars, so most of the data points here were clustered in the lower side of price. There were outliers in this dataset which could represent the customers without a strict budget buying luxury cars, which there were a few of in the dataset like the Catilac Escalade. However, the dataset has more data for the general consumer cars, so it makes sense that these customers are more price conscious. A takeaway from this plot for the car manufacturers is they should try to make their cars more affordable to appeal to price conscious people which is the majority.



My next scatter plot compares sales to fuel efficiency, and this plot surprised me a bit because I would think that more sales would come from fuel efficient cars. In the scatterplot there is a slight negative correlation when comparing fuel efficiency to sales. I had to think a bit about why this would be the case but when including some factors I found earlier, it makes more sense. First of all, this dataset is 6 years old, when gas prices were much lower and fuel efficient cars weren't as mainstream. It also cost the manufacturer more to include hybrid technology back then making most hybrid cars more expensive, and with what we found earlier, customers are looking for more affordable cars. Another thing to keep in mind too is what we saw earlier in the correlation matrix with large wheelbase cars (usually trucks and SUVs), were really influencing sales. Because trucks and SUVs seem to be getting a lot of sales, and those cars aren't very fuel efficient, it makes sense why there isn't much correlation between sales and fuel efficiency. This analysis made me realize that on a surface level some plots might not make sense so it's important to consider background information.



The next plot has 5 partial dependence plots that compare the value of each variable to the impact it has on sales (partial dependence). The first plot shows the curb weight of the car, and we can see that the sales of a car increase with curb weight. At a certain point after the sharp peak there is no data (since this dataset is only for consumer cars and not semi trucks). This makes sense with my earlier conclusion that trucks tend to get more sales. The fuel capacity chart shows us that up to a certain fuel capacity, cars get more sales, but then it's impact starts to plateau after a certain point around 30 (gallons). This shows us that fuel capacity is important for sales but after about 30 gallons (usually the fuel capacity for trucks) it starts to not matter as much. Next plot is fuel efficiency, where the data starts at what looks around 15mpg where its influence on sales are highest, then we see a drop around 25 mpg in sales, but then an increase after up to around 40 mpg then a plateau. This tells us that cars with not that much fuel efficiency like trucks are a big sales influence, cars with around 25 mpg don't have a sales impacting fuel efficiency, but then we see a rise at around 35 mpg leading to higher sales. So this tells us that fuel efficiency does matter after a certain point of gained mileage, and can increase sales after around 30 mpg. This can tell us that if a car manufacturer wants to appeal to people who want fuel efficiency, they should aim for about 30 mpg, before the plateau where it doesn't matter that much but after the slump of around 25mpg. The next plot is horsepower which we see gradually less pull on sales as it gets higher. Like I mentioned earlier, this dataset is mostly average consumer cars and car brands, so it makes more sense that the more horsepower, the less of an influence on sales because more horsepower means the car is expensive, and from our first plot, the cars people are buying are the cheaper ones in this data. The last plot shows a decline in sales the more expensive a car is which was also determined in the earlier scatterplot. It’s interesting though to see the jump from ~$25k to ~$30k. This tells us that the majority will try to buy the car at the lowest price, but the usual price point for the majority that isn’t trying to find the cheapest possible will usually spend about $30k. Companies should realize these 2 demographics and make cheap cars for the majority consumer and another car at about $30k for that demographic, and avoid pricing cars at around $25k since there isn’t much of a market for that price point.



This last boxplot compares price and sales to vehicle type, being either a regular car or passenger car (more seating and larger). We can see there’s definitely more sales for regular cars given they are much less expensive on average. Passenger cars seem to have a lot more range and outliers in their pricing which could be because some SUVs are priced for families while some are luxury cars. Even though cars sell a decent amount more, the outliers in the passenger cars might tell us that some passenger cars are very popular. With what we gathered earlier with most consumers wanting to spend more money, trucks or long wheelbase cars being more popular, a car company might have success in the passenger cars if they kept their car affordable.

Conclusion

Although the model for my data came with some challenges, I was able to use trial and error to get the best model I possibly could using ridge regression with cross validation. If I were to do this project again I would have probably used another, more up to date, dataset that also would work better for a model. I was under the assumption I would be able to improve the model after my first submission, and after many attempts I was able to lower the RMSE a bit but not as much as I hoped. Aside from the model, I found a lot of insights and contributing factors to car sales with my plots and I learned to use the context of findings in other plots to make more sense of other plots results.