Used Car Data Analysis

Kevin W

Contents

Introduction

Transportation is the life-blood of civilization. For many people, the two largest purchases in their lives will be a house and a car. However, not all cars are built equally and, from personal experience, many car salesmen will not eagerly disclose these inequalities. Thus, I will be investigating the used car market for my project, thinking from the perspective of a prospective buyer and from a used car dealership.

For this investigation, I used a dataset containing information of around 1000 unique car listings on Kijiji in the Greater Toronto Area.

Research Question

How do factors such as year, mileage, wheel configuration, and body type influence the price of used cars?

In particular, I want to find actionable insights that can benefit people who are considering purchasing a vehicle, whether they want to save money, want to find a vehicle with a low depreciation rate, or to understand what is within their budget, given a set of desired properties.

Methods

Data Source

My data was acquired as a `csv` from this github page. For each unique brand/model/year combination in the dataset, I scraped the 2023 market price of the car from MotorTrend.com using `Python`. Motortrend states that the market price is a "reasonable asking price by a dealership" for a vehicle with no defects and minimal wear.

Note that the market price was obtained in 2023, and may be lower than the listing prices obtained in 2019 due to depreciation. However, with prior knowledge that the depreciation rate should be fixed for different categories of cars and can be adjusted for, I believe relationships between these variables are still worth investigating.

Data Wrangling

First, I removed columns that I did not want to investigate. Most of these columns contain are difficult to investigate without further data collection, such as Carfax report links and vehicle identification numbers. I also chose not to use the "addresses" column. I want to find general insights that can apply to people living in different provinces, thus I decided against investigating the effect of location on listing prices.

Below are the variables of interest in the dataset:

Variable	Description
Brand	eg. Honda, Audi, etc.
Model	eg. Civic, R8, etc.
Model year	eg. 2001
Listing price	Price on Kijiji, listed by seller(CAD)
Market price	A "fair" asking price for a good-condition vehicle, from MotorTrend.com(CAD)
Mileage	Miles
Body Type	eg. Convertible, Sedan, Trucks, etc.
Wheel Configuration	eg. AWD, FWD, etc.

I then checked the validity of values in the remaining columns. I noticed many zero values for the listing price. This usually happens when a seller does not give a listing price(putting "Please Contact" on their listing) or gives a low listing price to gain priority when buyers sort listings by price. Either way, the zero values are not valid, and were removed.

I then checked for missing values. The only remaining column with missing values was the "Market Price" variable I scraped from Motortrend. Since market prices are scraped in 2023 and the dataset is from 2019, there is no reasonable way to impute the missing values in that column. Market price is an important variable in this investigation, so I decided to remove all rows with missing market prices. In total, I removed 196 rows.

I also noticed that some variables had blank values. I decided to map these blank values, found in the variables wheel configuration and body type to "Other," because I did not want to remove missing values. "Other" was already an existing category in the data, so I decided to cast all unknown values to that.

Finally, I added a new `price range` variable that indicated if the car mentioned in a listing had a market price below the 25th percentile, between the 25-75th percentile or above the 75th percentile, comparing to other recorded listings in the same year. This variable is intended to roughly indicate whether a vehicle is a "luxury vehicle" or not, and takes the values 'low,' 'medium,' and 'high.' I chose to compare with other vehicles of the same year, because prices may depreciate over time, thus comparing vehicles from different

years may not be a good indicator of whether a vehicle was in a high, low or medium price range when it was sold.

Results

Plots

I used plotting to investigate the data. Overall, I generated four figures, given in the "methods and results" section of the project webpage under the heading "Market Price Visualizations." Some of the plots presenting main results will be repeated below, but the interactive visualizations are given on the webpage and may be more informative.

Figure 1: Market vs Listing Price

Figure 1 checks my assumptions that market price is highly correlated by listing price, and that it may vary proportionally to the price.

The first plot shows that market price seems highly correlated to the listing price.

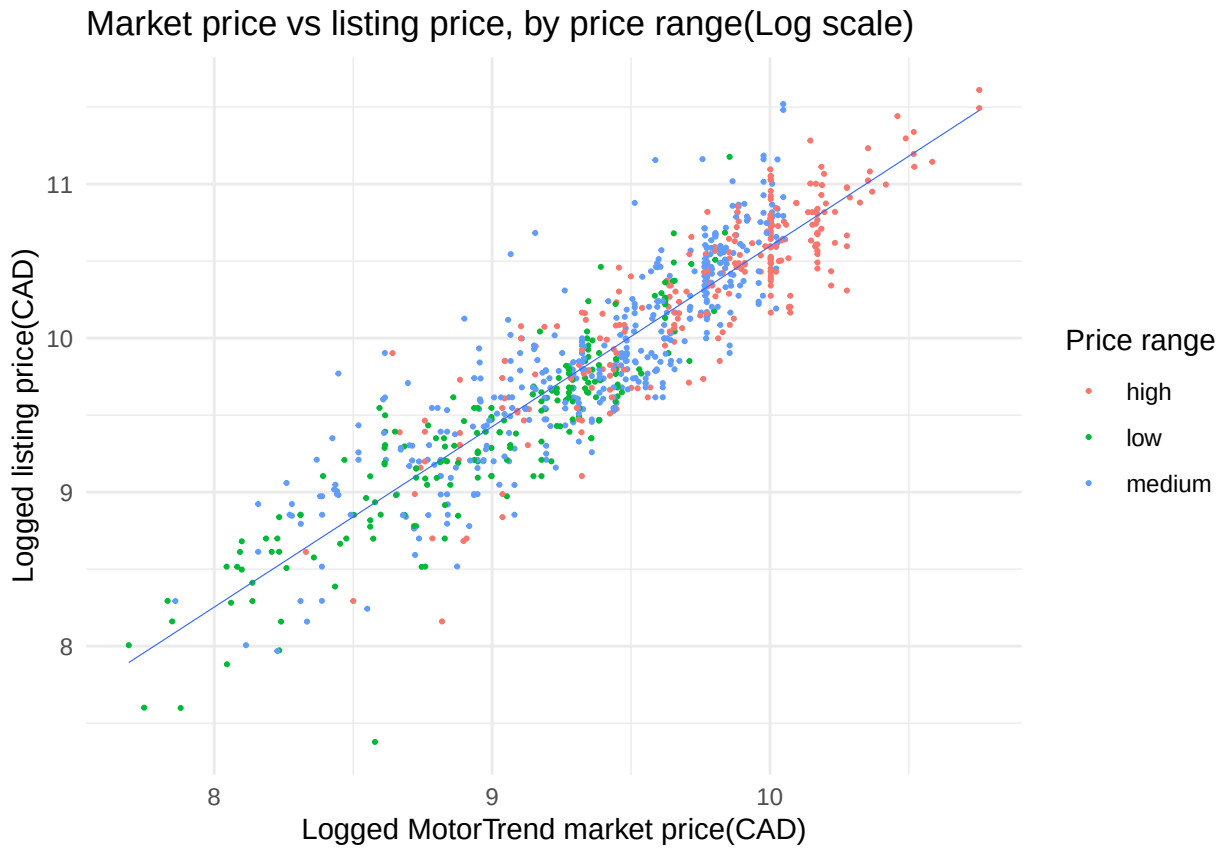
I fitted a line for listing price vs market price using least median squares regression, which is less influenced by outliers compared to least mean squares regression. The line has the equation

Listing price(2019) = -2663.14 + 1.8 * Market price(2023)

The model suggests the value of most used cars has almost halved since 2019, given the newer models that are being developed and the simple passing of time.

The cone shape of the plot suggests the variance in listing prices is associated with the market price of a car. It is common knowledge that factors affecting price such as depreciation and damage are all scaled to some base price of a new, accident-free vehicle(eg. an expensive vehicle would have expensive repairs, and the value would decay proportional to the original price)

To support the previous claim, we see that on the log scale, the plot has relatively constant variance across market price values. This suggests that the ratio of listing price to market price is somewhat constant across cars of all price ranges.



In conclusion, Figure 1 suggests that price changes in listings are proportional to each car's "initial price," given upon production. Car listing prices are thus correlated with estimated market price. Further factors must be considered if we want to explain remaining variance in the listing prices.

Figure 2: Logged market price by year and brand

Figure 2 allows us to visualize and compare the changes in price for different brands of cars over time. I chose to plot the market price variable, instead of the listing price, as it has less random influence from individual sellers. I also chose to plot market price on the log scales and fit linear lines, due to prior knowledge that listing prices decay exponentially. This setup models the exponential price decay.

Thus, for each brand, we fit a straight line $price = e^{\beta_0 + \beta_1 year}$, meaning the slope estimates roughly correspond to annual rate of change in price(the rate is estimated to be e^{β_1}.

Overall, the plots suggest that there isn't a significant difference in depreciation between different brands of cars. However, since within each brand we have observations for different models, the results are heavily affected by outliers and a lack of data points, and further investigation is necessary to confirm these findings.

The webpage contains the fitted slope coefficients for each line, that can be interpreted using the above description. For example, Porsche, Acura, Toyota and Honda are among the lowest slope estimates among

both cars and trucks, suggesting that they have the lowest annual price changes. Meanwhile, the models suggest that brands such as Cadillac, Audi, Dodge and BMW experience the highest annual changes.

This may be because brands such as Acura, Toyota and Honda are known to have low maintenance costs and encounter few mechanical problems in the car's lifespan, whereas brands such as Audi and BMW typically have high maintenance costs. However, the amount of noise and outliers in the data makes it difficult to draw definitive conclusions.

In conclusion, Figure 2 suggests that depreciation rate does not differ too much between brands, and consumers should feel free to choose brands that they enjoy. However, buyers may still want to consider the trend in a car's price over time before buying a vehicle.

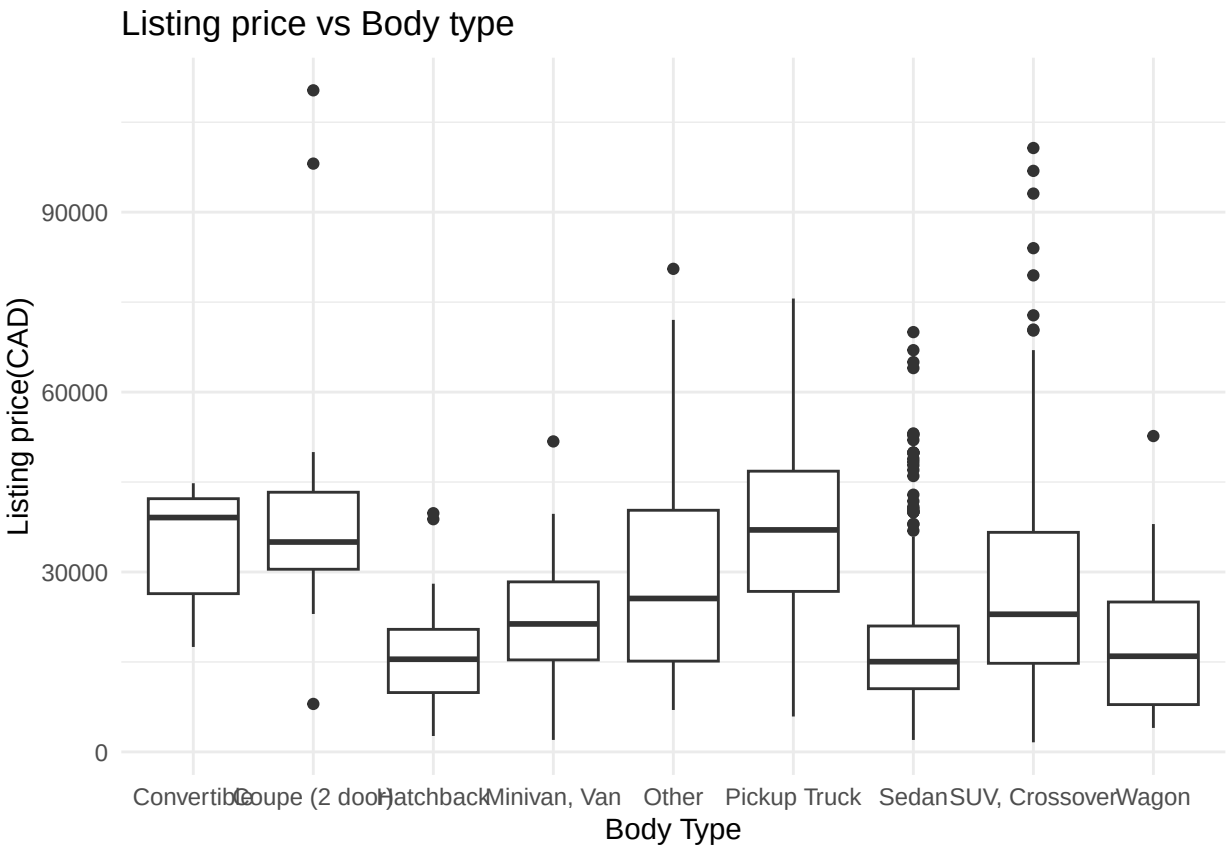
In addition, I fit a least median squares regression model to find the overall trend in logged market price over time, and got the equation:

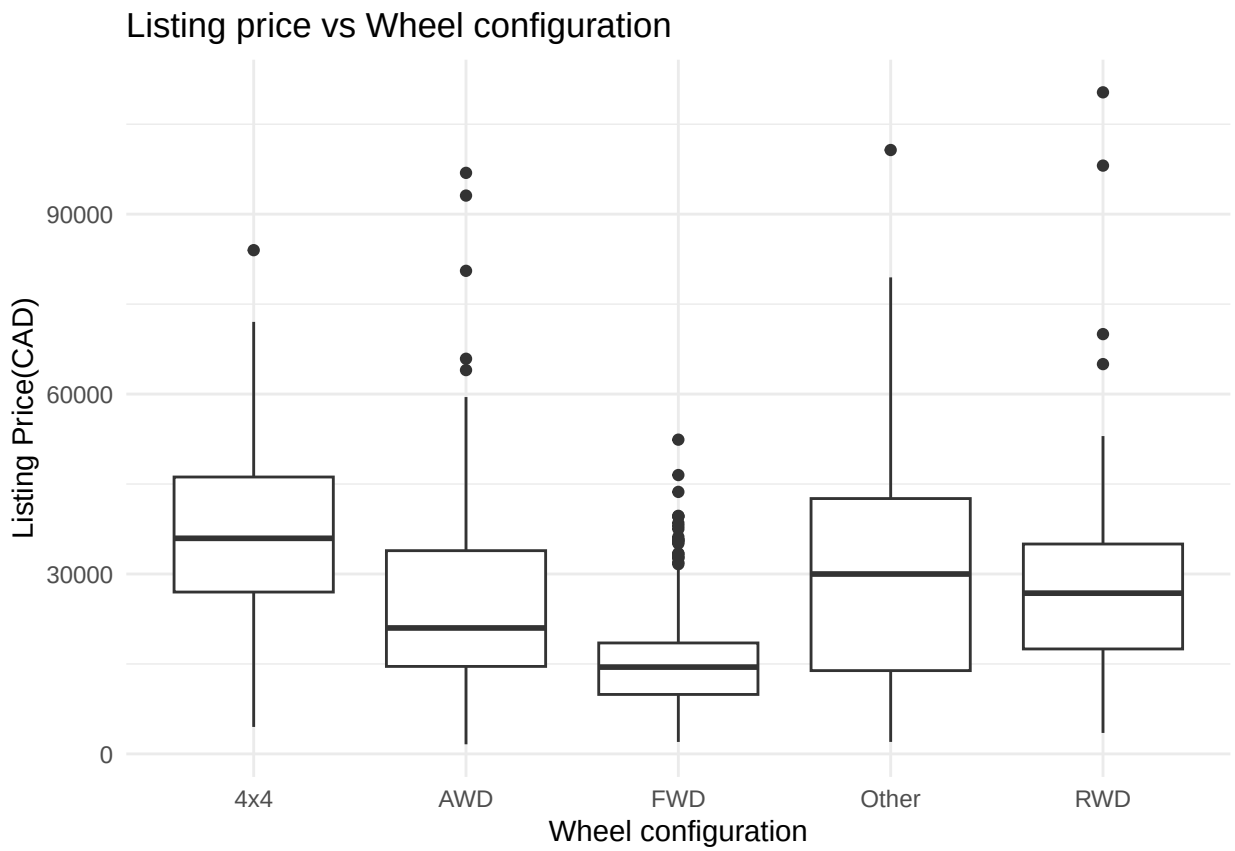
$$log(MarketPrice) = -252.6 + 0.146 * year \implies MarketPrice = exp(-252.6 + 0.146 * year)$$

This suggests that on average, a car's market value will be $e^{-0.146} = 0.864$ times that of a car of the same make and model, that is one year more recent. This may act as a helpful baseline depreciation rate when comparing individual makes and models.

Figure 3: Comparing other car factors

When considering a car purchase, knowing approximately what is available within a certain price range is important. Analyzing the price distributions of cars with different body types and wheel configurations can help us understand what cars are available at different budget points, a useful consideration when a specific function is required(eg. a truck for transporting equipment, or an all wheel drive vehicle for snowy weather).



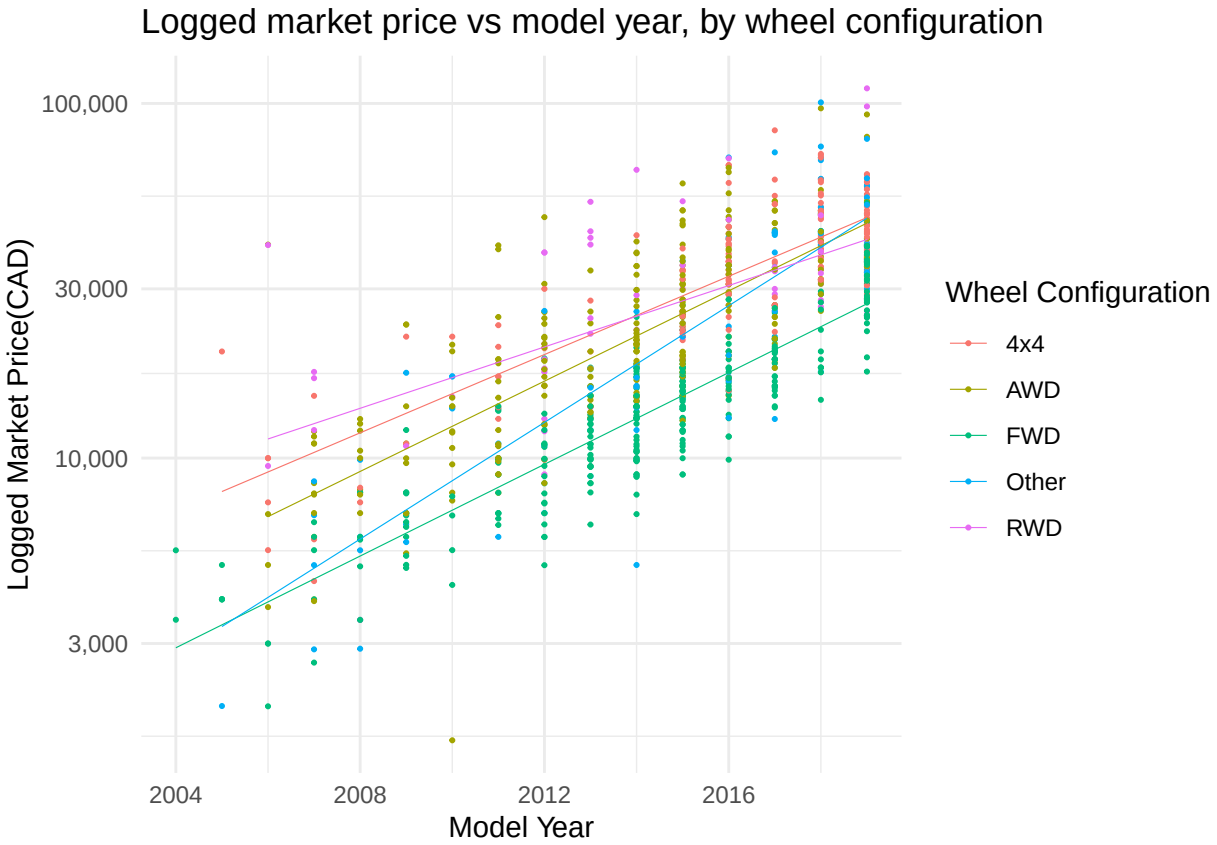


The visualizations suggest cars with different body types have varying listing price distributions. This makes intuitive sense, as larger vehicles will probably have higher prices compared to smaller cars. We see that pickup trucks, convertibles, and coupes have the highest price.

Overall, this suggests that body type may associated with pricing, and may be a good indicator for predicting car price. In addition to body type, wheel configuration also seems to be associated with price. We see 4x4 vehicles have the highest median price.

Figure 4: Comparing market price decay with miscellaneous factors

This visualization compares logged market price by year, separated by wheel configuration.



I fit least median squares regression lines for price vs year, separated by wheel configuration and body type.

Table 2: Fitted slope coefficients for log(listing price) vs year, by wheel configuration

Wheel Config	Slope Coefficient
RWD	0.0729573
4x4	0.1278840
FWD	0.1521631
AWD	0.1693435
Other	0.2051976

Table 3: Fitted slope coefficients for log(listing price) vs year, by body type

Body Type	Slope Coefficient
Coupe (2 door)	-0.0304588
Pickup Truck	0.1055533
Hatchback	0.1254380
Sedan	0.1362090
Minivan, Van	0.1496552
Other	0.1503306
Wagon	0.1616646
SUV, Crossover	0.1752411
Convertible	0.2169924

The plot, as well as the numerical statistics, suggests that trucks, that are usually 4x4 vehicles, and sports cars, which are usually RWD vehicles, observe lower decay rates than more "generic" vehicles such as cars and vans. We see that trucks and coupe(2 door) vehicles have the lowest slope estimates, and 4x4 and RWD vehicles have the lowest slope estimates among all wheel configurations.

This may be because trucks and luxury vehicles may have lower supply in the used car market, and thus customers are willing to pay higher prices. Also, customers may be seeking additional utility and status from their purchase.

Luxury vehicles and sports cars provide the customer with social status, additional to the utility of a car, thus causing customers to be willing to pay a higher price for old vehicles. Also, customers may require trucks to accomplish certain tasks. Since they cannot easily find another option to accomplish their goals, they are willing to pay higher prices for a used truck.

On the other hand, cars, SUVs and vans are easily replaceable, and have higher supply, and thus have a higher depreciation rate.

Overall, there is some evidence in support of our analysis in this section, but the size and noise in the dataset makes it difficult to draw any definitive conclusions.

Modelling

I fit multiple models that attempt to predict the listing price of a car given other variables in the dataset.

The models are meant to investigate how much of the variation in listing prices can be explained using the data I collected. Essentially, I want to evaluate how "useful" the data would be in predicting listing prices, using state-of-the-art prediction methods, such as random forests and gradient boosting. In addition, it helps us understand variable importance. The models are not created for predictive purposes. For future investigations, we would probably scrape present-day data and fit a new model if we wanted to make accurate predictions.

In this section, I was not really concerned with the interpretability of my models, as I believe conclusions given by multiple linear regression or decision trees are not really intuitive/helpful to prospective buyers, and more intuitive and useful observations have already been made using visualizations.

Linear Model

I initially considered a model using all predictors, as well as interactions between model year and brand, due to the results found in the visualizations earlier. The model attempts to predict log listing price, transforming the variable to satisfy constant variance assumptions.

The model includes market price as a predictor, which is strongly correlated with model year and mileage, so I chose to omit variables related to it, after some hypothesis testing.

I removed the mileage predictor because it is strongly correlated with year(Pearson's correlation = -0.752). Intuitively, older cars will most likely have higher miles. I also removed model year, as it was not significantly related to the response given the other variables.

I will omit the fitted coefficients of the model due to the amount of them, and because I want to mainly investigate the effectiveness of the model.

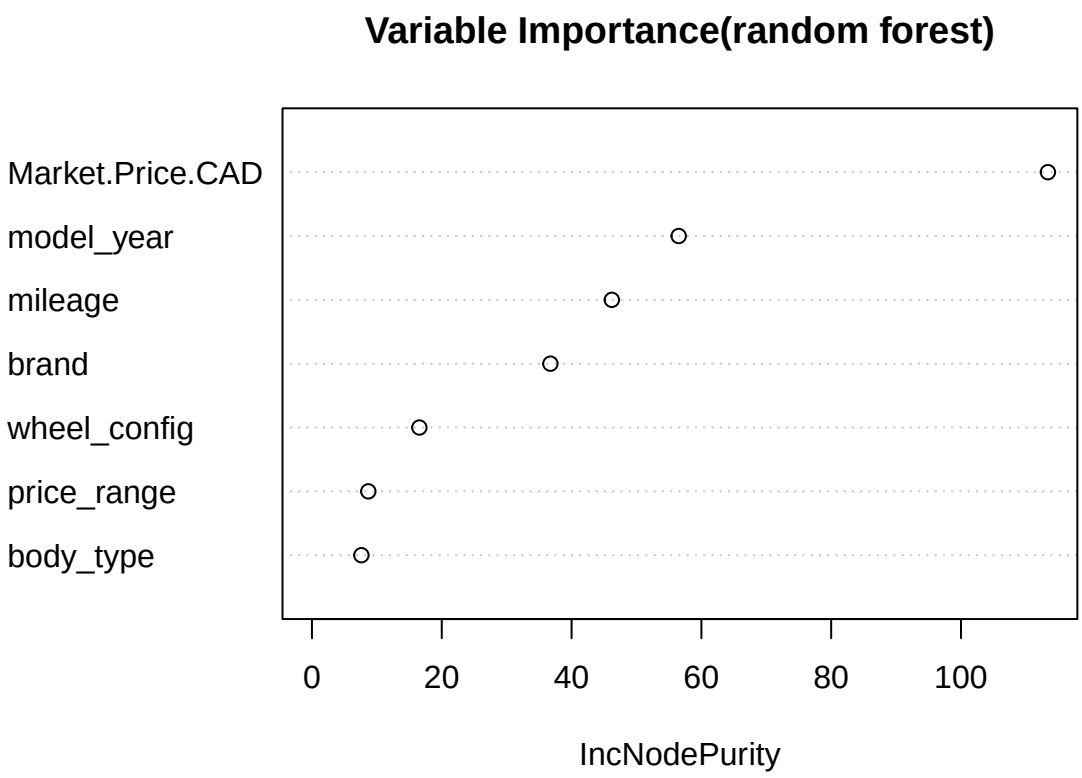
Table 4: metrics for the final(reduced) linear model

Item	Value
F-statistic p-value	$< 2.2e\text{-}16$
R-squared	0.8832

Random Forest

I fit a random forest model using logged listing price as the response, and brand, model year, mileage, body type, wheel configuration, market price and price range as predictors. I omitted the car's model as a predictor as there is not enough data on individual models for it to be used.

Below is the variable importance plot:

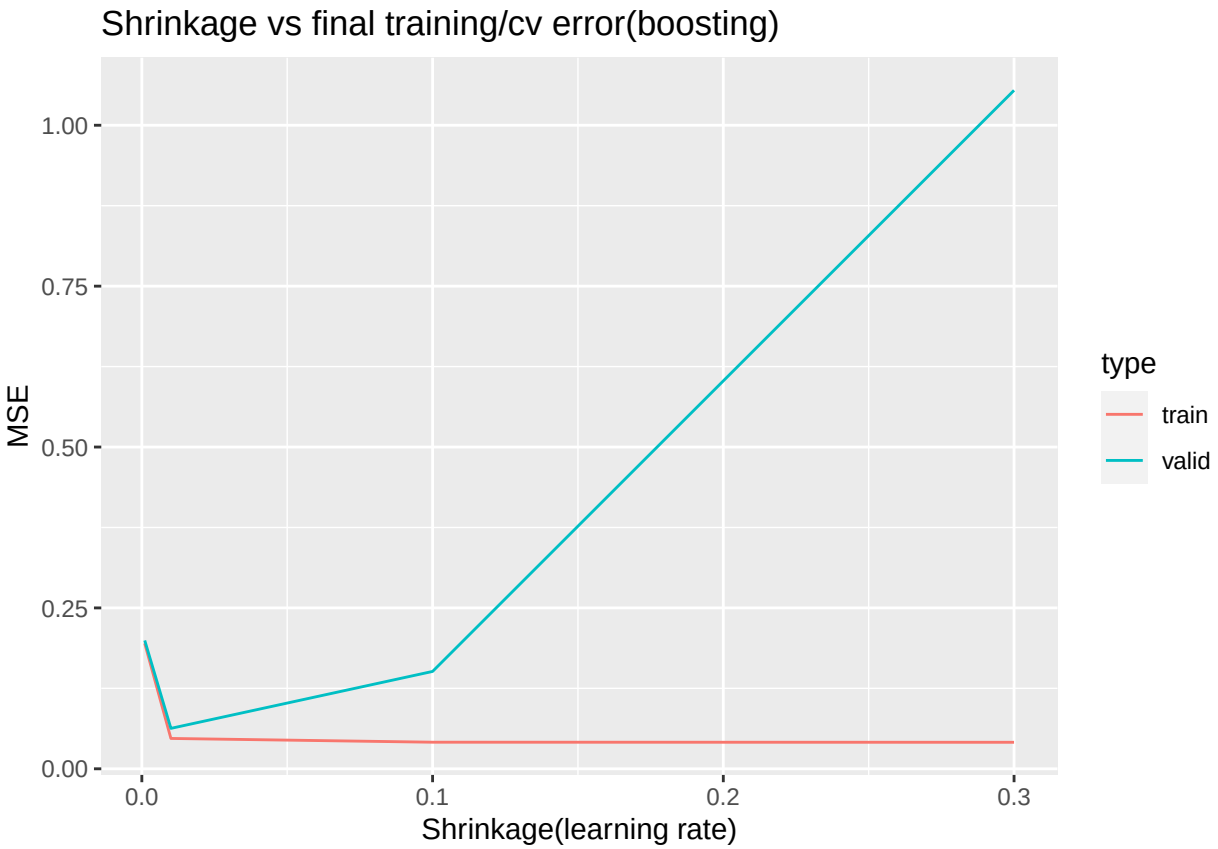


We see that market price is important when comparing to listing price, most likely as it was shown to be highly correlated. We then see that year, mileage and brand are important predictors of market price.

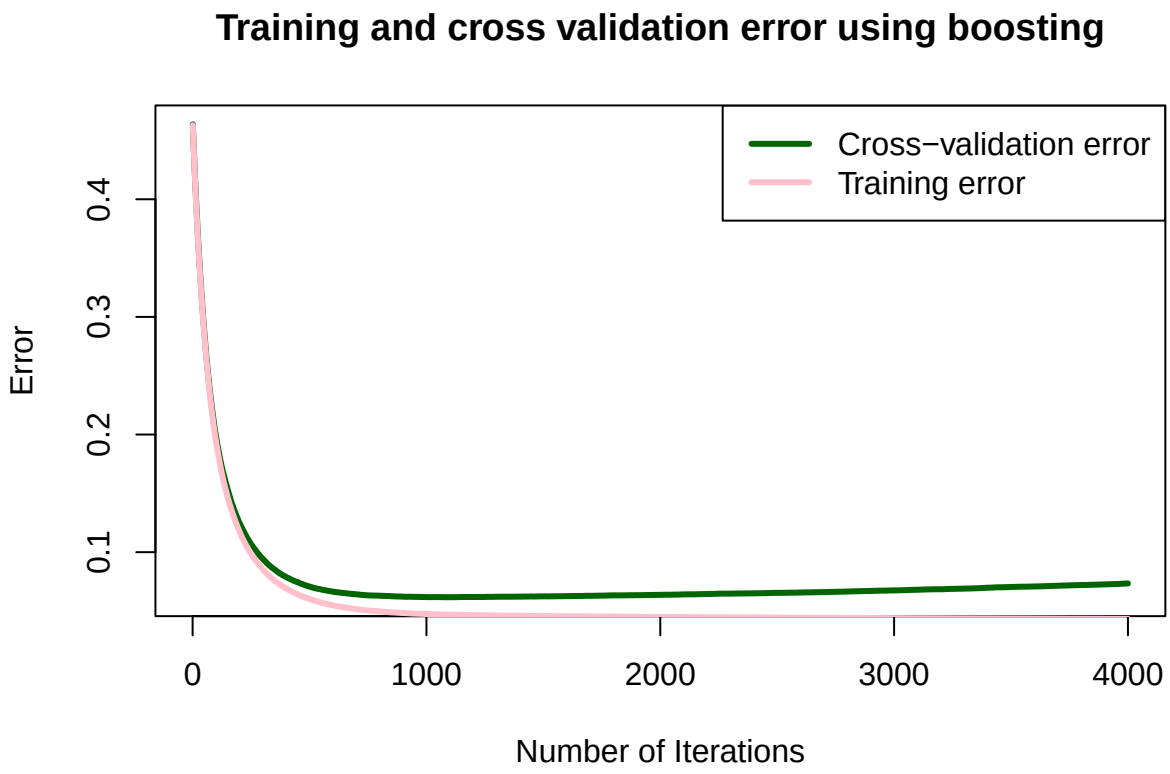
This suggests that the age and use of a car(which are related) and the price depreciation associated with time has a significant effect on price. Factors such as brand, wheel configuration and body type are less important. This may be because the price decay was shown to be exponential, and is thus significantly more impactful on price than the other variables, which may linearly or quadratically impact price.

Gradient Boosting Model

I then fit a gradient boosting model using the same formula as that used in the random forest. To tune the model, I measured the 5-fold cross validation score based on learning rates of 0.3, 0.1, 0.01 and 0.001.



This plot shows that the model seems to diverge with higher learning rates, and trains too slowly with a low learning rate. Overall, the learning rate of 0.01 seems to be optimal.



I then tried to fine-tune the number of iterations. It seems that the model converges in around 800 iterations before beginning to overfit.

Table 5: Variable relative influence in the boosting model

	Relative Influence
Market.Price.CAD	85.4850081
brand	6.2786269
mileage	4.6597031
wheel_config	1.8998624
model_year	0.9377146
body_type	0.7390849
price_range	0.0000000

In this model, we see that market price has a high relative influence, most likely for reasons previously discussed. However, we see that brand and mileage have a large influence, and model year has a low influence.

Model year and mileage are correlated, and are related to market price. Thus, it may not be important in the model due to collinearity. Intuitively, brand could be influential in the model as certain brands are known for selling certain types of vehicles(eg. BMW, Audi are known for luxury vehicles, while Ford, RAM are known for trucks and 4x4 vehicles).

Models summary

To compare the MSEs of these models, I constructed two baselines.

The first simply predicts the mean listing price in the training data as the listing price for all test observations(equivalent to fitting a linear model with no predictors).

The second is a simple linear regression model using only logged market price to predict logged listing price. This would be equivalent to consulting online resources such as MotorTrend.com when deciding the value of a car.

Below, we can compare the test MSEs of all the models.

Model	Test MSE
Mean	0.4766231
Simple LR	0.0806050
Multiple LR	0.0739132
Random Forest	0.0533401
Boosting	0.0602181

We see that simple linear regression offers a considerable decrease in MSE compared to the base mean prediction. Also, the multiple linear model offers little improvement in MSE over the simple linear model. This suggests market price is an important predictor, likely as it takes into account many of the other predictors already when it is calculated.

Advanced regression methods such as random forest regression and gradient boosting offer a sizeable improvement in performance over multiple linear regression. This suggests there may be nonlinear associations within the data.

Overall, the random forest gives the lowest test MSE. This may be because the random forest model is created to be robust to outliers. Given the small amount of training data used, this setup may be optimal compared to other models.

Conclusion

The visualizations and numerical analysis mostly suggest that most brands of cars decay at a similar exponential rate every year. In addition, we find that prices differ between types of cars and wheel configuration, and that certain types of cars such as trucks and sports cars may experience less price decay than normal cars due to the features they provide the buyer.

Finally, we see that state-of-the-art methods, such as random forests and gradient boosting, are able to make reasonably good predictions on the listing prices of vehicles, suggesting that the variables collected in this investigation could potentially be used for inference.

Limitations & Future directions

There are several limitations in this study. Firstly, the size of the dataset is small, and thus many conclusions about sparse variables such as brand do not have much supporting evidence and are prone to outliers. In addition, variables such as model could not be investigated as we lacked the data to do so, and thus more specific investigation could be done in the future.

In addition, the dataset is not recently collected, and thus I could only perform descriptive analysis. In the future, an updated dataset could be made, allowing more actionable conclusions to be drawn.

Finally, Kijiji listings may not be representative of all used car listings, and the listing prices of the cars may not accurately reflect the final sale price of each car. This discrepancy could be investigated in the future.

In future investigations, I have two general directions for this project. To make better inference, we could attempt to collect updated information to build a strong predictive model. To perform further analysis, we could attempt to understand how market price is calculated by collecting more information related to each car, and investigating relationships between those variables and the market price.