Using Statistical Learning Models to Predict Listed Prices of Used Cars

Eliza Malinova, Zhenghao Li, and Raushan Baizakova

Abstract

Constructing accurate statistical models to predict used car prices has been of high interest in numerous fields. Determining the resale value of a vehicle has been of a paramount importance for financial services, lenders and vehicle leasing services. This report examines number of distinct car specifications, thar are considered to affect vehicle prices the most. To construct an accurate statistical model aimed at predicting used vehicle prices, we applied four statistical techniques: Multiple Linear Regression Analysis, Lasso, Random Forest and Boosting. The predictions are then examined and compared to decide which model provides the best performance. The report concludes with analysis of particular vehicle specifications, specifically fuel type, to examine and approximate their effect on used vehicle prices.

Introduction

Currently, there are about 287.3 million registered cars in The United States. In addition, the dollar value of annual sales of cars is usually from 3 to 3.5 percent of GDP. The report aims at examining what car characteristics are the most important in predicting listed prices. However, the motivation to explore this topic is probably less academic, but more rooted in our experience of suspicious listed car prices found online. Therefore, if one has a well-trained price-predictive model to be referred to, it will be more assuring even when buying cars from a lemon market.

Since previous and current research has shown that those specifications like make, year, mileage, type, size, and other dimensions are probably important to predict prices, our analysis also follows this approach but uses different statistical learning models to figure out the complexity hidden behind those specifications. By doing so, it is more likely to get models performing better in predicting prices. However, car prices prediction is not an easy task, though a popular topic, which requires some expertise and special knowledge as well as more complete data. So our analysis is just a starting point from which many improvements can be done such as pre-processing data with a more professional understanding of cars, compiling more data across the nation, or adding information about both supply sides and demand sides of cars to adjust prices.

The rest of this report is organized as follows: Section 1 presents visual illustrations of the dataset, section 2 briefs the dataset and the methods/models used. Section 3 presents and discusses the main results of our models, and section 4 summarizes the main conclusions of our report. In the Appendix, some relatively less crucial but important tables and figures are listed to be viewed.

Section 1: Visual Illustration of the Dataset

The top panel of the figure 1.1. illustrates that the majority of the car types in the data are presented by Sport Utility Vehicles, sedans and pickup trucks. The bottom panel shows that each type has a relatively distinctive price range. For example, pickup trucks as a type have the average price of about 17 thousand

dollars, which is the highest among all types, whereas the average SUV costs 13 thousand dollars. Hatchbacks and wagons as a type have the relatively lowest average price of about 10 thousand dollars.

Figure 1.1: Top: Distribution of car types. Bottom: Boxplots of car prices by type

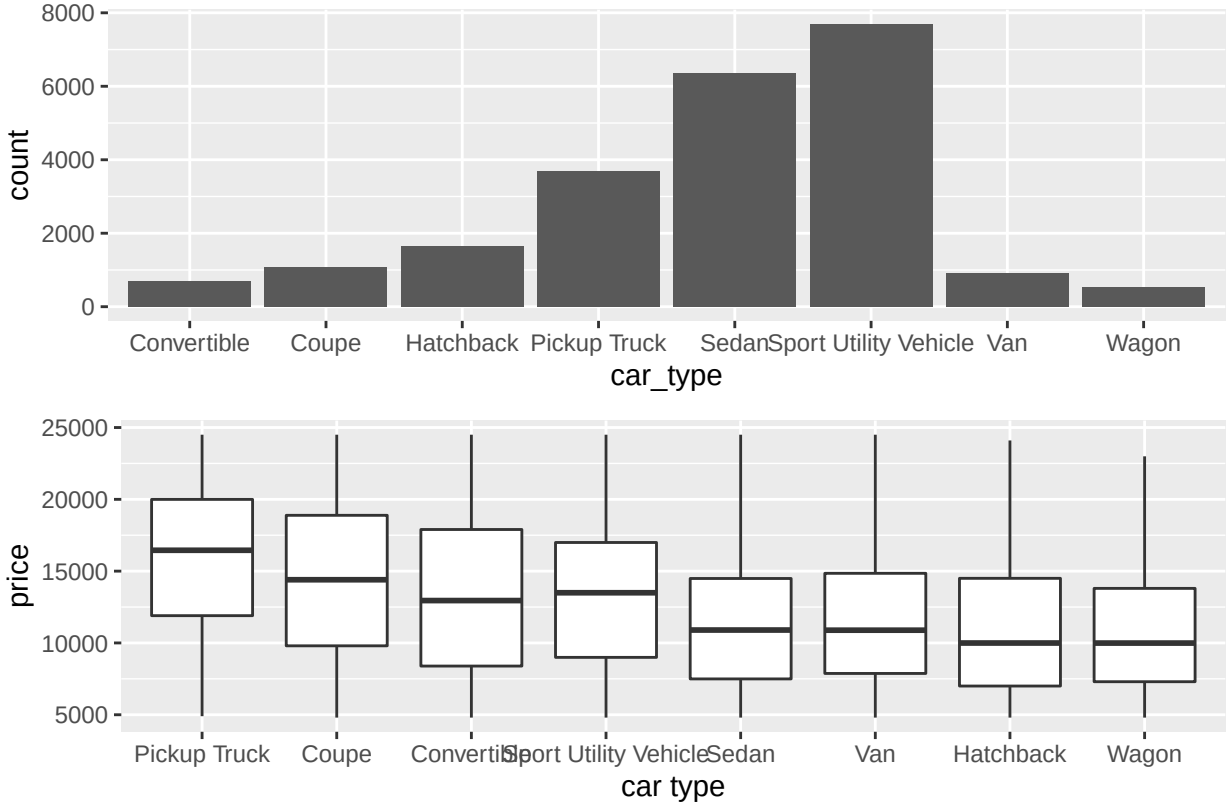


Table 1.1 in the appendix represents all the 47 car models in the dataset. One of the goals in the paper is to find how much the type of the engine, and the fuel it supports, affect the value of the car. This is the reason why the table illustrates the different types of engine fuels that the models have in our dataset. Afterwards, we visually illustrate the relationship between listed price, vehicle models and engine fuel types.

Results from the table also indicate the current situation in the car industry: even though electric and hybrid vehicles are becoming more popular, there is still a large prevalence of regular vehicles. The internal combustion engine vehicles that are powered mainly by gas prevail in the dataset used.

The vehicle models that also have both regular engine type fuel and hybrid/electric type are Audi, BMW, Buick, Cadillac, Chevrolet, Ford, GMC, Honda, Hyundai, Infiniti, Kia, Lexus, Lincoln, Mercedes-Benz, Mercury, Nissan, Porsche, Saturn, Smart, Subaru, Toyota, Volkswagen. However, we chose only these models that have the highest number of hybrid or electric vehicles to visually represent and analyze the price given engine type. The five models represented are BMW, Honda, Lexus, Nissan and Toyota.

Figure 1.2 represents a grid of five graphs to show the relationship between listed prices, vehicle models and engine types. The engine types here have been further grouped into Hybrid, Electric and Traditional. All engine fuel types beside hybrid and electric have been combined into Traditional to represent the regular type of vehicles. Even though Ford and Chevrolet have high amount of hybrid and electric vehicles, the regular type of engines exceed the hybrid and electric by so much that the visual representation is ambiguous, and we excluded these two models. The rest of the five vehicles models also have the potential to well represent the relationship between listed price and engine type since all five have high amount of electric and hybrid

vehicles.

Figure 1.2: Five Model Cars and Their Listed Prices By Odometer and Engine Type

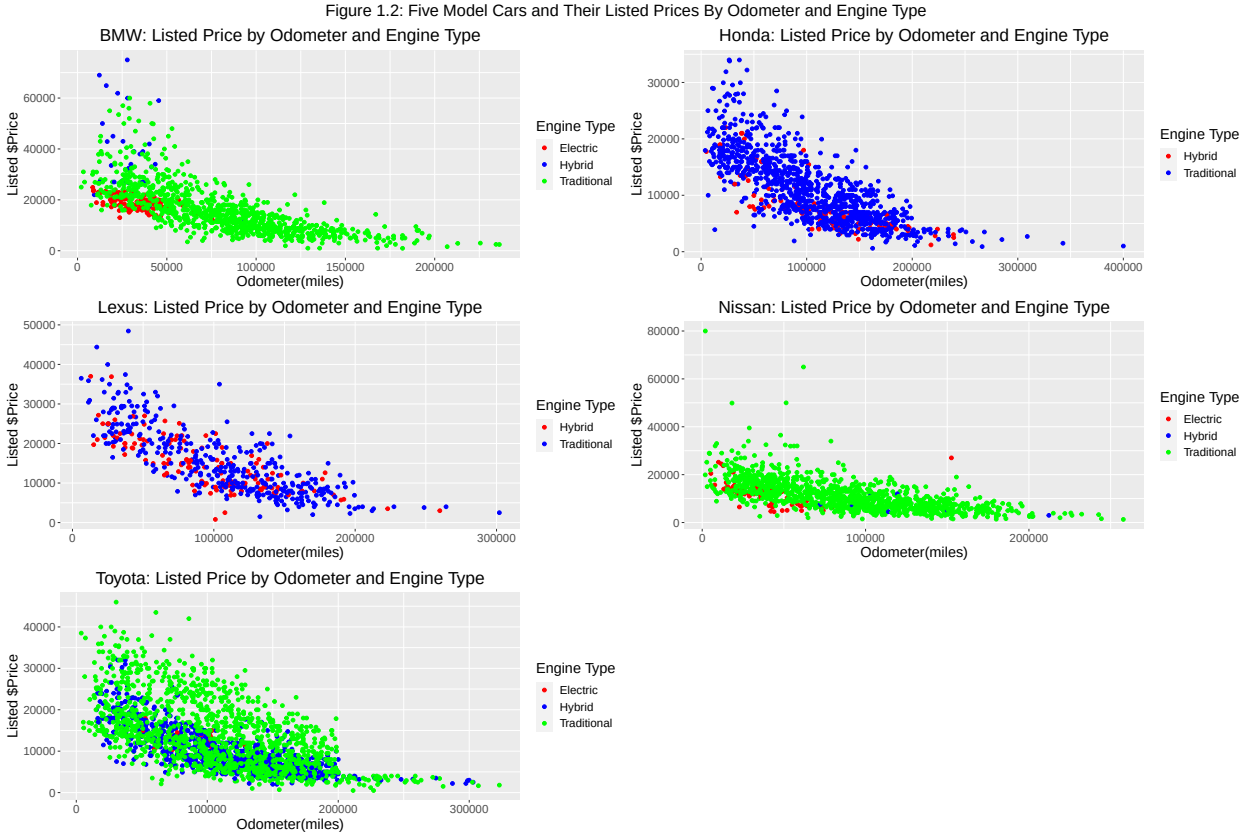


Figure 1.2 illustrates an expected relationship about the listed price and odometer: as the vehicles has more miles on it, the price increases, regardless of model type and engine type. However, the graphs do not show any distinctive relationship between the prices of traditional vehicles compared to hybrid and electric vehicles. Considering Honda, Lexus, and Nissan, we can conclude whether hybrid and electric vehicles have higher price on average by concentrating only in a particular range of odometer values. On the other hand, if we focus on the lower range of odometer: between 0 and 50000 for BMW, we can see that we can expect higher price of BMW hybrid vehicles compared to electric and traditional. Looking at the Toyota vehicles, we see the opposite relationship than what is illustrated at the BMW graph.

Analyzing these five graphs above, we keep in mind that we do not account for any other features of the vehicles other than odometer and engine type. The listed prices of cars are affected by many other vehicle characteristics such as car model, color, size, year, and other attributes.

Section 2: Data Methodology

The dataset used for estimate is an assembled one. Originally, it was a Craigslist dataset which was posted on the Kaggle platform (https://www.kaggle.com/austinreese/craigslist-carstrucks-data). However, this original dataset has some caveats: many missing values, too few specifications of cars, and some values are inaccurate. To overcome these limitations, we extracted only valid vin numbers, which serve as the unique identity of a car, from the original dataset, and then used these vin numbers to get relatively full specifications of cars

from CarsXE (https://api.carsxe.com/), a database of cars with API access. After removing duplications, a subsample in a size of approximately 24,000 cars was drawn to do our analysis.

The variables used in our analysis are summarized below: Price: listed price of a car on January, 2020 Make: brand of the vehicle car_type: types of cars by body styles car_size: vehicle size classes made_in: country manufactured Fuel_Engine: type of fuel used for car fuel_capacity: fuel tank capacity measure in US gallons engine_size: The size of an engine measured in cubic centimetres (cc) engine_cylinders: number of engine cylinders transmission_type: type of transmission transmission.speed: number of speeds of transmission drivetrain_ad2: type of a drivetrain car_door: number of doors in car curb_weight: total mass of a car without passengers and cargo overall_height: the vertical dimension from the highest point on the car to ground overall_length: the maximum dimension measured longitudinally between the foremost point and the rearmost point on the vehicle overall_width: the maximum dimension measured between the widest points on the vehicle, excluding exterior mirrors wheelbase_length: the dimension measured longitudinally between front and rear wheel centerlines standard_seating: number of designated seating positions odometer: distance travelled by vehicle measured in miles mileage_range: indicator for the range of distance travelled by vehicle in miles city_mileage: miles travelled per gallon in the city highway_mileage: miles travelled per gallon in the highways avg.mileage: average of city mileage and highway mileage.

Section 3: Statistical Models

The first model is the baseline model, a hand-built linear regression model including some interactions. The second model presented is Lasso model which includes all interactions. The main aim of Lasso model is for variable selection, to select these features with nonzero estimates of coefficients. This gives us a better understanding of variable interactions and their effects on the car prices. The other two methods to be used are two tree-based: boosting and random forest, which tend to capture potential interactions and non-linearity better as well as fit with a dataset with many categorical variables. At last all above methods are measured by cross validation and are compared by their error rates (averaged RMSE).

Section 3.1: Linear Regression Models

Before hand-building a linear regression model, a stepwise selection method was applied to our data. Without any interactions, the stepwise selection just returned a linear model including all features. Once interactions included in the stepwise selection, it became time-consuming and cumbersome in running. Moreover, through a simple 90:10 training and testing splititng, the performance of hand-built model was better than the setpwise selection model. So, we used the hand-built model as the baseline model. In addition, since including a few interactions improved the model performance, we expected the interactions would play an important role in prediction. However, instead of checking features and their interactions in the baseline model, we did that in lasso regression, which deals with interactions much more efficiently.

Table 3.2.1 in the appendix gives hand-built linear model's estimated coefficients of the various car features in predicting listed prices that are statistically significant at 0.1%. Each coefficient shows a ceteris paribus effect of every feature on car prices. For example, as might be expected, coefficients on year show that as cars get older, the estimates on year decrease holding other car features constant. Estimates on makes reveal interesting facts about some cars: make Porcshe is associated with about $13500 price increase from the base level of make Acura holding other features constant, whereas make of Ram - with the lowest price decrease of about $2200. In general, it can be noticed that cars' listed prices tend to increase with the fuel capacity, engine size and cylinders, and as cars mileage travelled increases, their listed prices tend to decrease non-linearly. Having large-sized vehicle independently on the type tends to have very negative effect on the

listed price, whereas having midsize sedans and wagons have positive effects. Interestingly, in can be noted, despite being relatively old vehicles, certain types like hatchback, sedan and wagon tend to have higher prices, holding other features constant. In the case of wagons, we can even see that older ones tend to have higher prices. Lastly, consistent with the graphs before, our estimates show that on average, having electric car is associated with a $6000 increase in price holding other vehicle features constant.

Section 3.2: Lasso Model

To perform Lasso regression, the data is scaled first. The model is ran based on the "poisson" distribution which not only resembled the non-parametric distribution of price but also gave us a slightly lower error rate compared to "guassian". In addition, following the default rule, we chose lambda.1se instead of the minimum lambda in prediction, since such lambda is likely to lead to a more parsimonious model. And by comparing the results of lambda.1se and lambda.min, no significant difference in estimated coefficients or error rates is presented between them.

By performing Lasso Regression, it can be observed that interactions do have impact on predicted used vehicle prices. Without including any interactions, the Lasso gave 64 nonzero coefficient estimates out of 101 coefficients, implying about two thirds of our variables had predictive power. However, after including interaction, only 259 out of 4117 variables/interactions had nonzero estimates, and the coefficient estimates of most single variables turned to zeros. The Table 3.2.1 listed top 20 variables/interactions impacting prices (ordered by the absolute values of estimates and excluded the intercept).

Table 3.2.1 Most Important Lasso Selection Variables	
Variable/Interaction	Coefficient
makePorsche:car_typeCoupe	0.4197
makePorsche:car_typeSport Utility Vehicle	-0.3355
made_inItaly:Fuel_EnginePremium Unleaded Gas	-0.2539
makeVolkswagen:car_sizeLarge	0.2426
makePorsche:car_door4	-0.2410
makeJeep:transmission.typeManual	0.2347
made_inUnited Kingdom:transmission.speedHigh Transmission Speed	0.2040
makeMitsubishi:car_typeSedan	0.1994
makeMercedes-Benz:transmission.speedHigh Transmission Speed	0.1883
makeChevrolet:car_sizeLong Bed	-0.1798
makePorsche:Fuel_EngineRegular Unleaded Gas	0.1756
makeFord:car_sizeLong Bed	0.1697
makeSmart:car_typeCoupe	-0.1633
makeJaguar:car_typeCoupe	0.1565
makeJeep:car_sizeMidsize	0.1511
Fuel_EngineHybrid:transmission.speedHigh Transmission Speed	0.1479
makePorsche:transmission.typeAutomatic	0.1342
makePorsche:standard_seating	0.1302
made_inUnited Kingdom:Fuel_EngineFlex-Fuel	0.1279
made_inMexico:transmission.speedHigh Transmission Speed	-0.1271

In Table 3.2.1, all are interactions, which strongly implied us the existence of complicatedly interactive and non-linear relations among those variables. Moreover, after including interactions, the error rates of Lasso improved a lot, as shown by Table 3.2.2 below.

Table 3.2.2: Error Rates of Lasso With and Without Interactions	x
Without Interaction	3884.171
With Interaction	3322.125

Sections 3.3 and 3.4 present two decision tree based models: Random Forest and Boosting. Considering Lasso variable selection ,previous knowledge and reasoning, total of 23 independent variables have been used to train the models. The distinct car specifications used as explanatory variables aremake, vehicle type, size, country manufactured, engine fuel, fuel capacity, city and highway mileage, steering type, wheelbase length, seating, transmission type, types of drivetrain, odometer, year manufactured, engine size, engine cylinders, and so on. The vehicle prices are as of January 2020 as well as all other car attributes including odometer mileage.

Section 3.3: Random Forest

The third statistical model performed is Random Forest. The decision trees built from bootstrapped training sets are 200 and a random sample of 4 explanatory variables are chosen as split candidates from the total number of predictors. The use of averaging the predictions of each of the resulting regression trees is designed to improve the predictive accuracy and control over-fitting. Since each split uses only 1 of these 4 variables, a sample of 4 predictors is taken at each split. In this way, Random Forest also overcomes the problem of highly correlated predictions from the 300 regression trees.

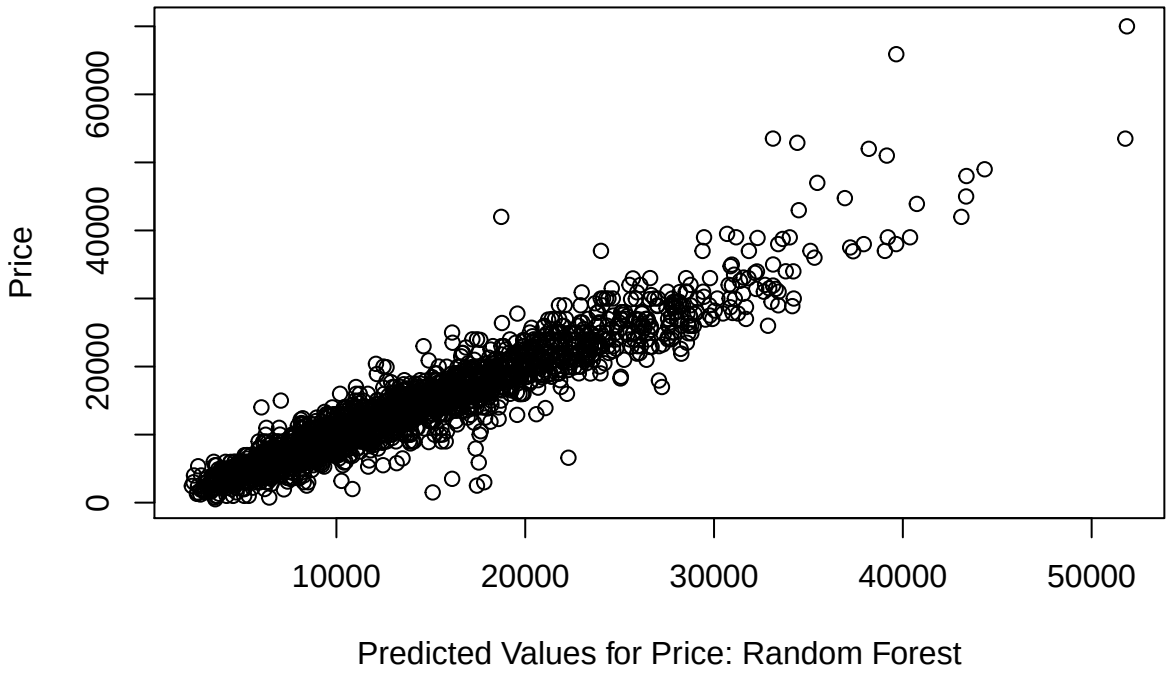
The data has been split in 90% train set and 10% test set. The RMSE result on average is below 3000 and the percent of variables explained is 88.3%. Given the RMSE, the Random Forest model can be considered of sufficient accuracy.

Table 3.3.1 illustrates the variable importance of each predictor. Running Random Forest, it can be recorded the amount by which Mean Squared Error (MSE) increases if the variable of interest is omitted from the model. In addition, the variable importance analysis shows the increase in node purity that results from splits over that variable. The increase in node purity reported is averaged over all trees. In this Random Forest model, the most important variables, based on percent increased MSE, are odometer, year manufactured and make. Considering increase in node purity, the most important variables are again odometer and year manufactured.

Table 3.3.1 Variable Importance in the Random Forest Model		
rn	%IncMSE	IncNodePurity
make	48.905937	94466535303
car_type	18.468822	46588280173
car_size	10.300440	15408977903
made_in	21.012515	22413737524
Fuel_Engine	29.650738	40609075498
fuel_capacity	9.886353	71744169302
highway_mileage	19.581779	52876799145
city_mileage	17.545911	44288034502
engine_size	17.666775	38704552066
wheelbase_length	19.809522	76768481061
standard_seating	19.390872	14110552417
transmission.type	15.789555	6793347956
transmission.speed	18.483115	32065777156
drivetrain_ad	22.821817	78839684705
drivetrain_ad2	12.256067	8688365883
car_door	11.657016	6360199324
curb_weight	20.989395	79867310960
odometer	55.074829	300011489007
year_factor	46.052940	309349376219
engine_cylinders	15.683840	25775936273

Figure 1.3 illustrates the prediction accuracy of the Random Forest model. The y axis represents the actual price of the test set and x axis represents the predicted price. It can be interpreted that most of the predicted values are close to the actual values of the test set.

Figure 1.3: Comparison between Random Forest Predicted Values of P and Actual Price Prices



Applying the Random Forest model, it can be examined how on average price will change if cars fall in different categories such as models, engine, and fuel type. Given the difficulty of interpreting results from Random Forest, the method we apply here and on the rest of prediction analysis is to predict prices on all cars that fall under a specific category or multiple categories and compute the average of these predicted prices. The limitation of this approach is that we do not account for all the variables, hence, conclusions of the exact effect of a specific category cannot be determined. A solution to this caveat can be to specify as many as possible variables such as model, year, fuel and engine type and range of odometer.

For instance, the average predicted price of a car that has an engine that supports regular unleaded gas is approximately $11887 while the average predicted price of a car that is Electric is approximately $14408. This is a very rough approximations since important car specifications among vehicles are different.

Being more specific in categories, if a vehicle is a 2013 Chevrolet model, using regular unleaded gas, then the average predicted price is approximately $9074, while a 2013 Chevrolet electric model has a price on average of $11098. In this scenario the rest of car attributes such as odometer and size still vary among cars.

Section 3.4: Boosting

The fourth statistical method used to predict prices is Boosting. The aim is to improve the prediction accuracy by using a method which, unlike Random Forest, grows decision trees sequentially. In this way each tree is fit using information from the previously grown trees. Both Random Forest and Boosting involve combining many decision trees, however, in Boosting each fitted tree can be thought of as "a modified version" of the previous tree. Each of the decision trees fitted in Boosting is usually small, but that contributes to the model's ability to improve the predicted value of the output by adding a new crushed tree each time.

The parameters chosen for boosting are as follows:

1) Number of trees: 500
2) Shrinkage Parameter: 0.2
3) Interaction Depth: 4

Table 3.4.1 presents the relative influence of each explanatory variable used in Boosting. Relative influence is a measure that indicates the relative importance of each variable used in training the model. The most important variables in the Boosting model are the manufactured year and odometer. This result coincides with Random Forest where the same variables were considered as the most important.

Table 3.4.1 Variable Relative Influence in Boosting	
var	rel.inf
year_factor	24.8961949
odometer	20.0927903
make	10.1760295
overall_width	9.4027653
drivetrain_ad	9.2383353
highway_mileage	3.7226198
city_mileage	3.6450222
overall_height	3.2816622
fuel_capacity	3.1773098
curb_weight	2.2404040
engine_cylinders	2.1612245
engine_size	2.1525728
wheelbase_length	1.3562452
Fuel_Engine	1.2121387
overall_length	0.9550451
car_type	0.8032286
made_in	0.5605983
X	0.2571783
transmission.speed	0.2124093
standard_seating	0.1635448
transmission.type	0.1065213
car_size	0.0994103
car_door	0.0867494
drivetrain_ad2	0.0000000

Section 4: Conclusions

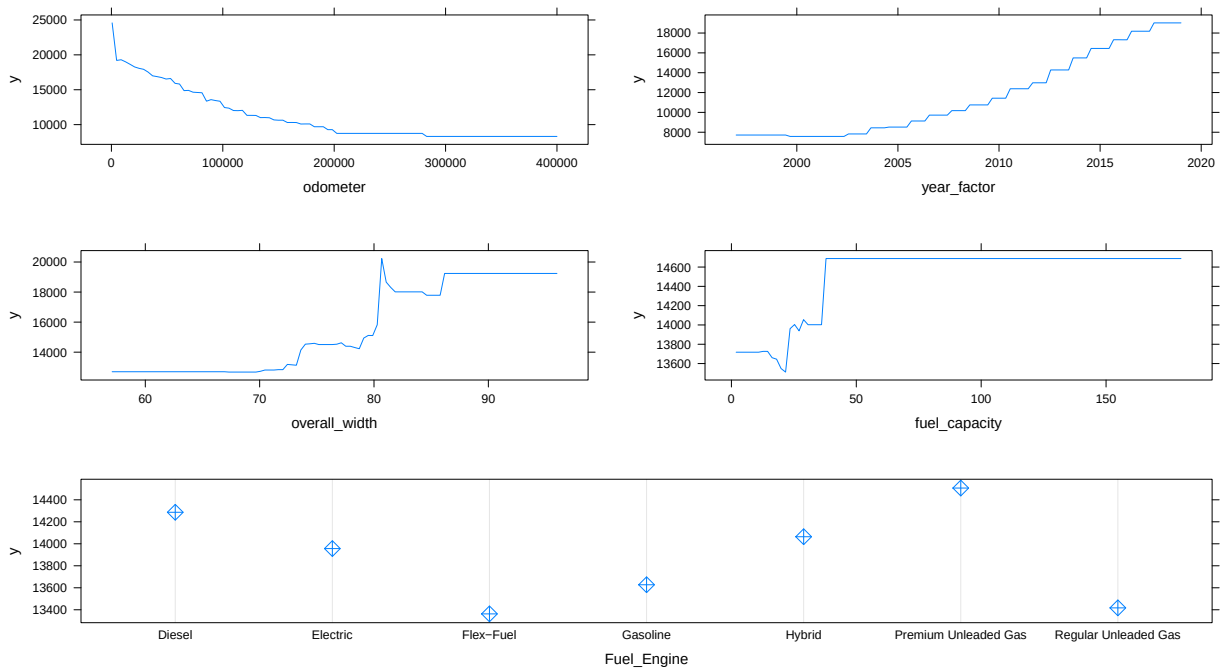
The error rates (averaged RMSE) of four methods in 10 - fold cross validations are presented in the Table 4.1 below.

According to the error rates, tree-based models: Random Forest and Boosting perform better on average than both Lasso and the Linear Model. Also, Random Forest performs on average slightly better than Boosting. So, we will mainly use the estimates from random forest as well as boosting to summarize the effects of features on prices. However, we will also refer to Lasso because Lasso not only showed us an improvement over the linear model, but also has the advantage that incurring less computation burden than two tree-based models.

```
##        Linear         Lasso      Boosting Random_Forest
##      4367.550      3334.921      2840.928      2857.449
```

Both Random Forest and Boosting produce similar results regarding variable importance. The following variables: odometer, year manufactured, make, overall_width. Fuel_engine and fuel_capacity are reported as most important. The partial effects graph of these variables (except make, which is shown by Table 4.1) are shown in Figure 1.4. The importance of of fuel_engine type and make is more subtle than expected. Examining the decision tree based model, it can be observed that odometer, year, overall_width, and fuel_capacity have non-linear effects on prices, among which increase in odometer will negatively affect price while increase in the other three will positively affect price.

Figure 1.4: Partial Effects of Top 5 Important Variables



Observing the statistical models' results, we can generally see cars of some fuel_engine types, such as Flex-Fuel and Regular Unleaded Gas, have lower average prices. But since cars are almost different in every other specification, for cars in specific categories, the effects of fuel_engine types may be different. For make, the same argument can be applied. At last, for fuel_capacity, there exists an upper bound above which variations in fuel capacity do not change the partial effect.

Table 4.1: Predicted Prices for Each Make	
make	average prices
Hummer	22226.05
Porsche	20594.59
Land Rover	19237.31
Lexus	17366.83
Pontiac	16375.78
Saab	16242.86
MINI	16201.55
Honda	16148.57
Acura	16023.52
Subaru	15906.38
Mercedes-Benz	15883.81
Toyota	15752.08
Saturn	15723.52
Scion	15711.98
Suzuki	15624.18
Mercury	15470.86
BMW	15056.95
Volvo	15045.34
Oldsmobile	14940.48
Lincoln	14919.44
Audi	14874.92
Cadillac	14841.02
Infiniti	14632.33
Smart	14242.93
Jeep	14151.45
GMC	14104.00
FIAT	13917.56
Mitsubishi	13864.60
Kia	13829.52
Jaguar	13797.51
Mazda	13716.02
Buick	13119.19
Nissan	13016.68
Ram	12986.46
Volkswagen	12967.18
Chevrolet	12918.24
Hyundai	12847.07
Ford	12717.28
Dodge	12329.24
Chrysler	12144.80

Following our previous method of approximating predicted prices given specific car attributes, we aim at observing how the different type of engine fuel affects the listed price of used vehicles. Table illustrates approximations of the average predicted used vehicle prices given the cars are regular unleaded gas and electric, accounting for odometer readings. The approximations of the predicted prices are constructed by the Random Forest model as described in Section 3.3. As it can be seen, one could expect that Regular Unleaded Gas cars will be cheaper than Electric cars if the car has an odometer reading over 40000 and year manufactured 2017. However, the opposite is observed in the case of 2017 vehicles that have odometer readings between 10000 and 40000. Accounting for car makes and Hybrid vehicles do not change the conclusions of the Random Forest model predictions. The average predicted prices of more environmentally friendly cars are higher than

regular cars in the odometer range above 40000 but lower in the odometer range between 10000 and 40000.

The above conclusions are rough approximations since we are not accounting for any other variables. However, as Random Forest and Boosting have reported, we account for the most important variables: make, manufactured year and odometer readings. Hence, we can expect similar prices to prevail in actual online listings.

```
##                                                          Average Predicted Prices
## Gas with odometer b/w 10000 and 40000                                       20897
## Electric with odometer b/w 10000 and 40000                                  17715
## Regular Unleaded with odometer over 40000                                   15844
## Electric with odometer over 40000                                           20591
## Gas Chevrolet with odometer b/w 10000 and 40000                               NaN
## Electric/Hybrid Chevrolet with odometer b/w 10000 and 40000                 16762
## Regular Gas Chevrolet with odometer over 40000                              14815
## Electric/Hybrid Chevrolet with odometer over 40000                          11647
```