Audi Used Car Price Analysis

Ashley Tian, Chenyu Wang, Rachel Liu & Xiao Hu

2021-12-05

INTRODUCTION

Motivation

When purchasing a used car from a dealership or a private party, it is important to know what factors may impact on the car's true remaining value. Customers with a budget constraint, such as students and low-income individuals, are vulnerable to unexpected problems when they buy a second-hand car. Technically, if customers may afford to own a new car sold by the manufacturer, they could be reasonably guaranteed that it has not been in an accident or tampered with. But a second-hand vehicle could have been used in any number of ways, and just because it does not seem to wear and tear does not mean there is not a problem. Moreover, customers also need to be concerned about the non-standardized price in the used car market. Theoretically, a car with more frequent use and shorter life could lead to a lower value. However, even if we would assume used cars are sold under good conditions, our data shows some fluctuation in prices of used cars registered from year to year. Hence, it is of our interest to analyze what variables may play a major role in affecting the selling price of a used car of a popular brand. In addition, we will further use the best fitted model to try to use the variables to predict the specific type of used car price using parametric and non-parametric methods.

Source of Data

All visualizations and code produced by the sample data of "Audi used car listings" by Mysar Ahmad Bhat, a Kaggle expert, are open access under the Creative Commons by license. The raw dataset contains 9 variables of 10668 Audi used cars registered from 1997 to 2020 with prices ranging from 1490 to 145000 Euros. For the purpose of the report, we perform an analysis on the linear relationship of the car price given 3 categorical variables and 5 quantitative variables.

Variables of Interest

In this report, the response variable is the price (in Euros) of an Audi used car. We consider that the price depends linearly on eight types of explanatory variables including the vehicle's model type, year of registration, transmission type, mileage, engine fuel type, road tax, miles per gallon and engine size (in litres). For example, among these covariates, the price of engine fuel (i.e., petrol and diesel) is likely an external factor that can linearly affect used car values because the vehicle models are rarely hybrid but powered by petrol or diesel.

EXPLORATORY DATA ANALYSIS

Numerical Variable

There are 6 numerical variables: year, price (in Euros), millage, tax, mpg (mile per gallon) and engine size (in liters). Price is the response variable of interest while the other five are explanatory variables.

The mean, median, minimum and maximum of all numerical variables are shown in Table 1. As we can see, the data was collected from 1997 to 2020. The prices of audi used cars in this data range from 1490 to 145000 Euros, with average price approximately 22897 Euros. In addition, there are unreasonable data including 0 mile per gallon and 0 engine size probably due to data collection error.

Table 1: Data Summary

	mean	min	median	max
year	2017.1	1997.0	2017.0	2020.0
price	22896.7	1490.0	20200.0	145000.0
mileage	24827.2	1.0	19000.0	323000.0
tax	126.0	0.0	145.0	580.0
mpg	50.8	18.9	49.6	188.3
engineSize	1.9	0.0	2.0	6.3

The distributions of numerical variables are not normal distributions (Figure 1). The year's distribution skewed to the left, while the rest of variables' skewed to the right. Most of our samples come from recent years (after 2015), and majority of the used audi cars have price lower than 50,000 dollars with the rest of them relatively more expensive. We might need to transform or standardize those variables, in order to meet modeling assumptions in the further analysis.

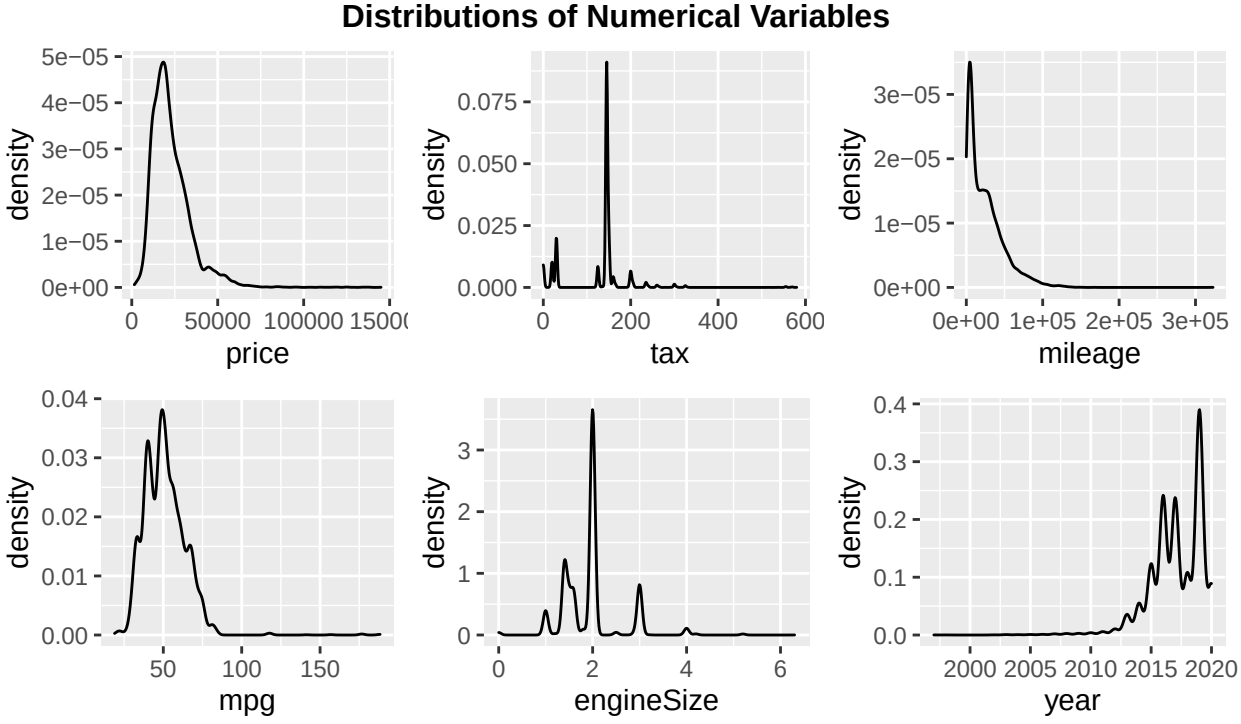

Figure 1

From the correlation heat plot (Figure 2), we can see price has strongest correlation with mpg and weakest correlation with tax. For collinearity issues, mileage and year have relatively higher correlation coefficient (r = -0.79) as well as the correlation coefficient between mpg and tax (r = -0.64). Ridge regression model might help to deal with the multicollinearity here.

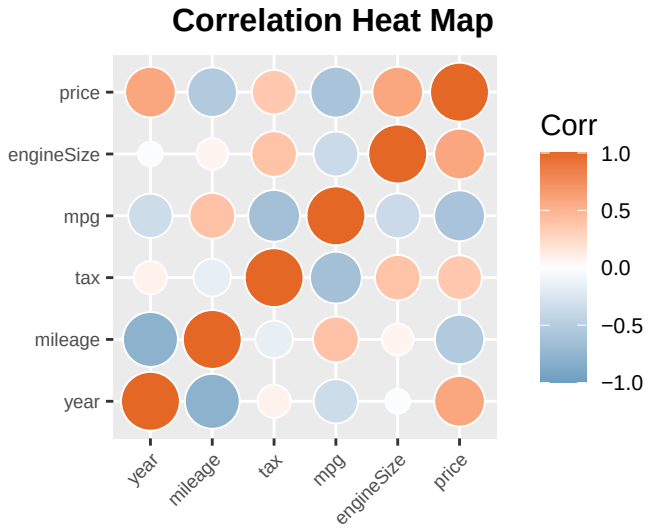

Figure 2

If we look at the scatter plot of tax and price (Figure 3), it also implies weak correlation relationship between these two variables (r = 0.36). What's more, results from scatter plots (Figure 3) indicated moderate positive correlation between price with registration year and engine size (r = 0.59 for both), while there is moderate negative correlation between price with distance used (mileage) and miles per galoon (r = -0.54 & -0.6 respectively).

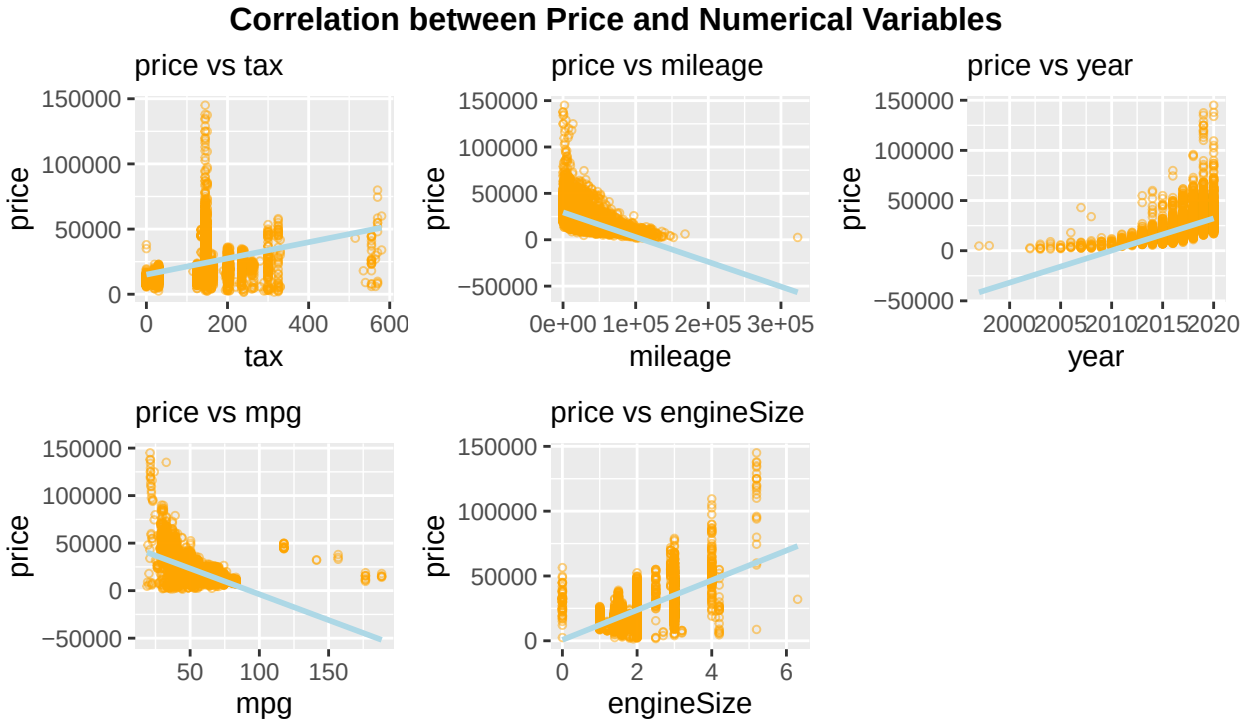

Figure 3

Categorical Variable

There are 3 categorical variables that could potentially affect the price of Audi used cars, which are model, transmission and fuelType.

There are 26 models in total. As shown in Figure 4, the Distribution of Model shows that the numbers of cars in each model differ a lot. A3 is the most popular model, A1, A4 and Q3 are also relatively popular, while models in R and S are not really popular. Also, the plot of Price against Model indicates that Audi used cars in some models have generally higher prices than the cars in other models. For example, model R8 is possibly more expensive than most of the other models. Therefore, the model is a possible factor for price.

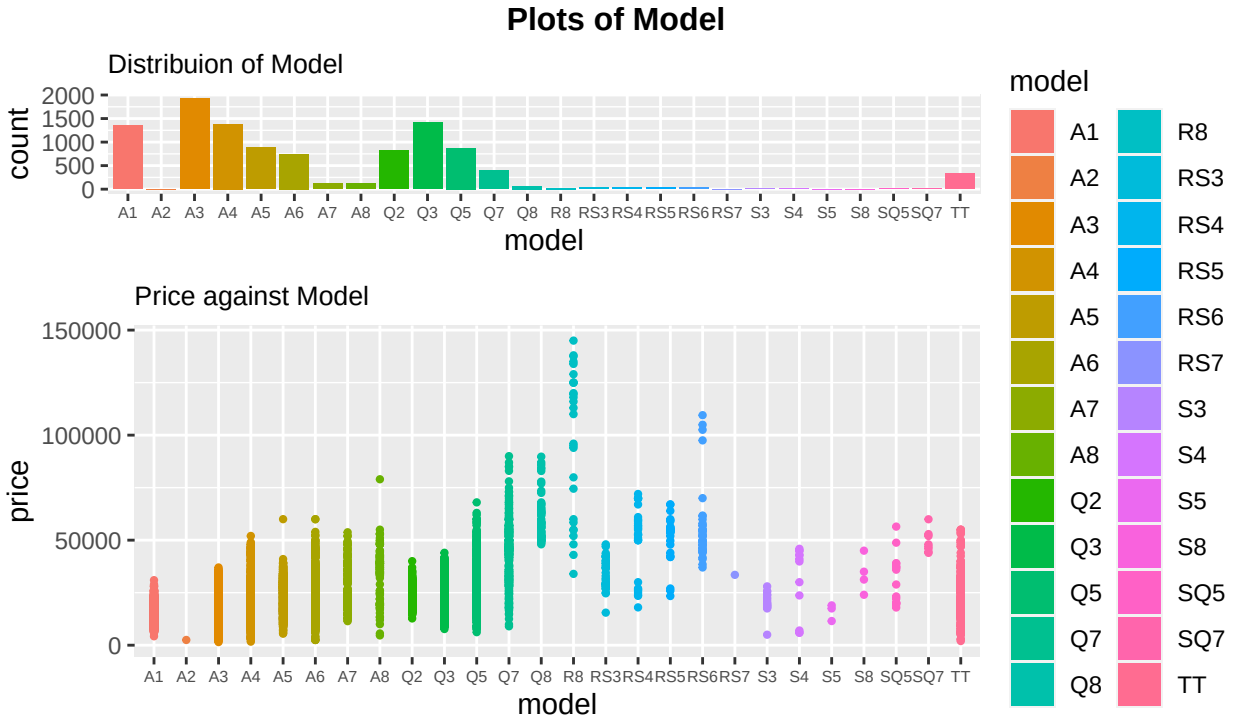

Figure 4

There are 3 kinds of transmission, automatic, manual and semi-auto. As seen in Figure 5, the Distribution of Transmission shows that the number of cars with manual transmission is the highest while the cars with automatic transmission are the lowest. Also, the plot of Price against Transmission indicates that cars with automatic and semi-Auto transmissions have wider price range and higher price than the cars with manual transmissions. Therefore, transmission is also a potential factor for price.

There are 3 types of fuel, diesel, hybrid and petrol. As shown in Figure 6, the Distribution of Fuel Type shows that most of the used cars use diesel or petrol while only a few used cars use hybrid fuel. Also, the plot of Price against Fuel Type indicates that the cars using petrol have the largest price range from 0 to 150000 Euros, while cars using hybrid fuel have the smallest price range from 0 to 50000 Euros. Therefore, fuel type could probably affect the price of Audi used cars.

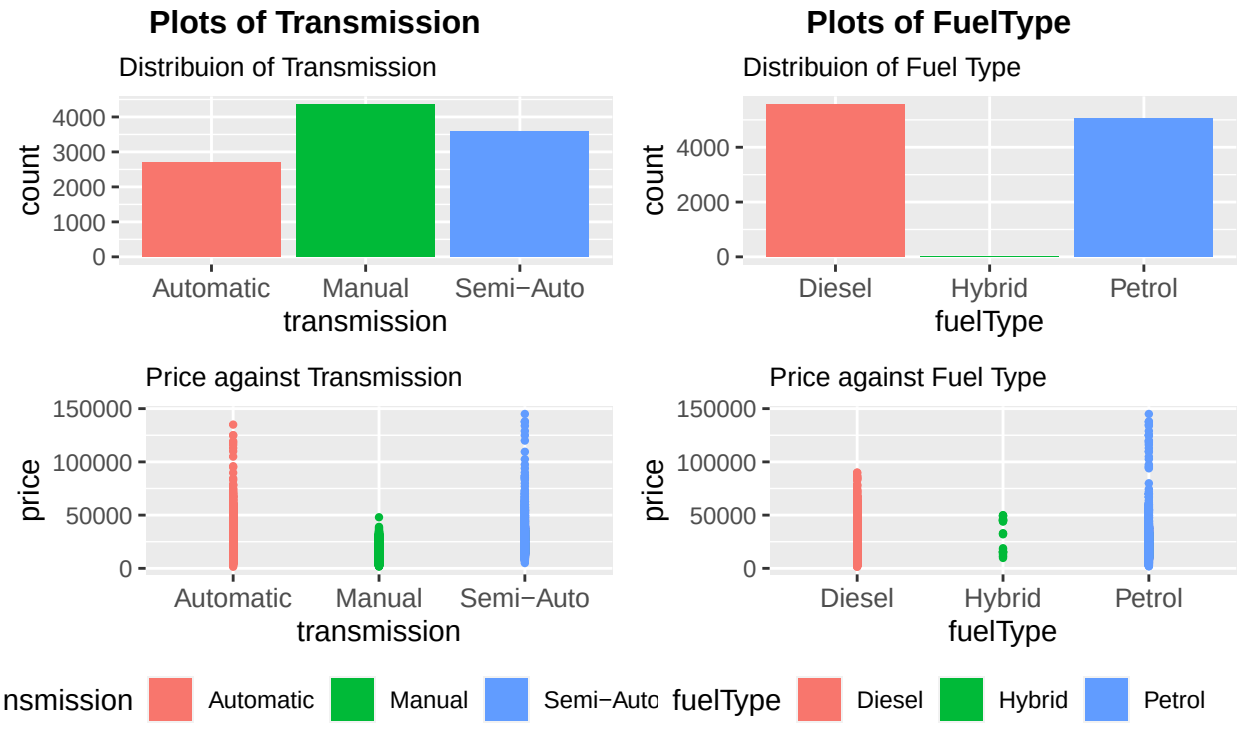

Figure 5

Figure 6

RESULTS AND ANALYSIS

Parametric Analysis

To explore the how the exploratory variables affect the price of audi used cars, we could do regression models and examine the coefficients of each variable. From EDA, we found that all exploratory variables have some influence on the response variable price, so we will use all these variables to fit the models. Moreover, in order to reduce the mean squared errors and fit the better model, we will apply ridge and LASSO regression.

We will first fit the following two models: ridge regression and LASSO regression, and then select the better one by computing and comparing their mean squared errors.

Ridge regression

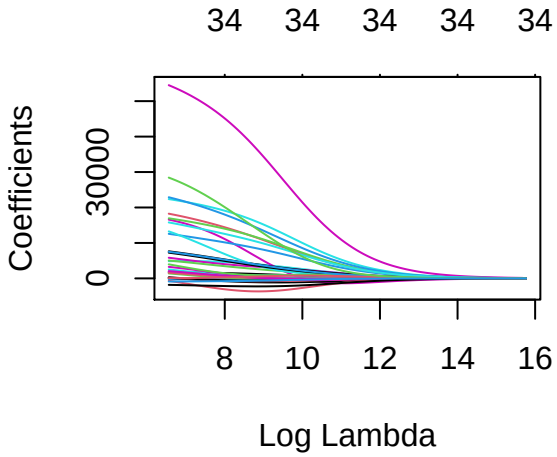

Figure 7(a)

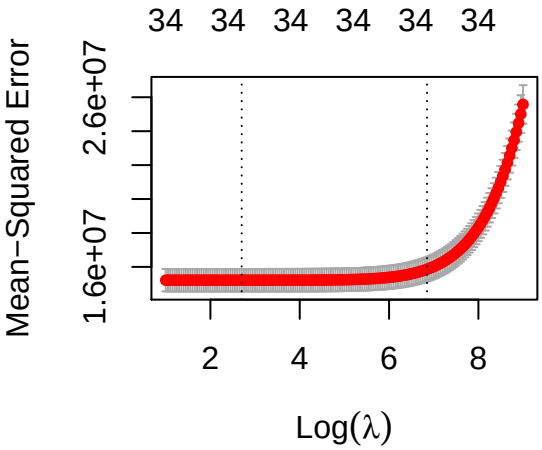

Figure 7(b)

As shown in Figure 7, the values of coefficients of numerical variables in the ridge regression model decrease as lambda increases. It is reasonable under the definition of ridge regression, that as the limitation increases, the coefficients shrink more. In addition, the minimum mean squared error is found at lambda equal to 14.88.

Use lambda = 14.88, the coefficients of the ridge regression model fitted are shown as following: (The coefficients of categorical variables are omitted.)

The mean square estimation error is 1.4849459×10^7 using this ridge regression model.

LASSO regression

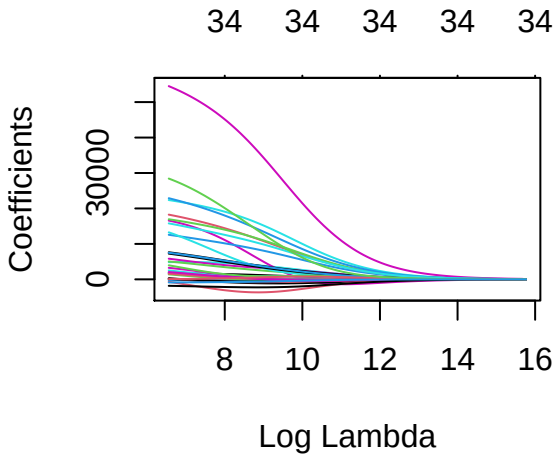

Figure 8(a)

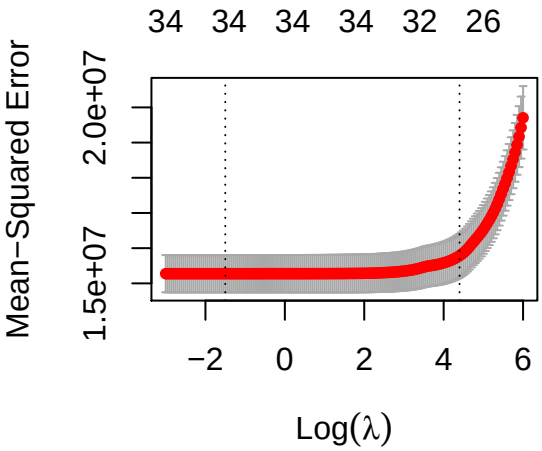

Figure 8(b)

As shown in Figure 8, the values of coefficients of numerical variables in the LASSO regression model decrease as lambda increases. It is reasonable under the definition of LASSO regression, that as the limitation increases, the coefficients shrink more. In addition, the minimum mean squared error is found at lambda equal to 0.223.

The mean square estimation error is 1.4849027×10^7 using this LASSO regression model.

Compare the two models

Here are the root mean squared errors for the two models fitted above:

Table 2: RMSE of Different Models

rmse.ridge	rmse.lasso
3853.5	3853.444

As shown in Table 2, the root mean squared error from LASSO regression model is the lowest. Therefore, we will choose LASSO regression model to estimate the price of audi used cars.

Use lambda = 0.223, the coefficients of the LASSO regression model fitted are shown as following: (The coefficients of categorical variables are omitted.)

Table 3: Coefficients in LASSO Regression Model

	Coefficients		Coefficients		Coefficients
(Intercept)	-3637459.458	modelA8	8205.926	modelS3	5022.794
audi.year	1819.215	modelQ2	1406.954	modelS4	9516.068
audi.mileage	-0.079	modelQ3	2886.989	modelS5	2053.811
audi.tax	-28.983	modelQ5	6776.597	modelS8	7383.441
audi.mpg	-288.195	modelQ7	15091.828	modelSQ5	9733.359
audi.engineSize	4446.233	modelQ8	24999.395	modelSQ7	19168.499
modelA2	19066.347	modelR8	60786.15	modelTT	3519.465
modelA3	1278.191	modelRS3	9849.966	transmissionManual	-1532.948
modelA4	1617.65	modelRS4	21444.727	transmissionSemi.Auto	96.054
modelA5	2982.167	modelRS5	19526.693	fuelTypeHybrid	33392.264
modelA6	3530.019	modelRS6	26996.961	fuelTypePetrol	-908.205
modelA7	4621.636	modelRS7	18969.512		

According to the coefficients shown in Table 3, variables year and engine size have positive relationships with price. While the other three numerical variables mile age, tax and mile per gallon have negative relationships with price. We could notice that the price of audi used car increases as year increases, which means that the newer cars are generally more expensive. Also, as mile age, tax and mile per gallon increase, the price will decrease. Additionally, used cars with larger engines are probably more expensive. In addition, it could be found that models Q8, R8, RS4 and RS6 are generally more expensive than the other models. Moreover, for the transmission type, manual cars are cheaper than semi.automatic and automatic cars. Finally, for fuel type, cars using hybrid fuel are more expensive than cars using Petrol and Diesel.

Non-parametic Analysis

From EDA section, the distribution of price with long right tail (Figure X) indicates that there are some second-hand Audi cars charged at a very high price. In the analysis, we set our budget equal to the medium price ($20200) and identified the 'expensive' cars as those with prices higher than 20,200 dollars. We aim to predict what kind of cars tend to be 'expensive'. This is a classification problem, therefore we decided to use the random forest model to predict the outcome and tune the hyperparameters through grid search.

In the random forest model, since previous study (Probst and Boulesteix, 2017) has shown that the biggest performance gain can be achieved with over 100 trees, we set number of trees as a sufficiently large number (400). In the study conducted by Probst et al. (2018), tuning the parameter `mtry` provides the biggest average improvement of the AUC, while tune of the other parameters doesn't have obvious effect. Therefore, for simplicity and time-saving, we only tune the parameter `mtry`, which is the number of randomly selected in each tree. In terms of grid search, each axis of the grid is an algorithm parameter, and points in the grid are specific combinations of parameters. Because we are only tuning one parameter, the grid search is a linear search through the vector of number of possible predictors used in the model (from 1 to 8). For each value of `mtry`, we utilized 10 fold cross validation with 3 repeats to get the accuracy of each model. From the plot of fitting results (Figure 9) after repeated cross validation, our best model should include all eight predictors, and the accuracy is 0.94, which means if we have all information about the 8 variables, then we are able to predict whether the price of this Audi car is over our budget or not with 94% accuracy rate.

The confusion matrix indicates that this random forest model have even performance in predicting expensive cars and not expensive cars for the training set, and the error of out-of-bag samples is 0.0583052.

```
##      predicted
## true    0    1
##    0 5028  308
##    1  314 5018
```

Our variable importance ranking (Figure 10) is based on the decrease of Gini impurity when a variable is chosen to split a node. From the variable importance plot, `year`,`mileage` and `mpg` are the three most useful variables in random forest model. Therefore, if we plan to purchase or sell a used Audi car, the top 3 factors that we need to consider should be the car's registration year, mileage and miles per gallon.

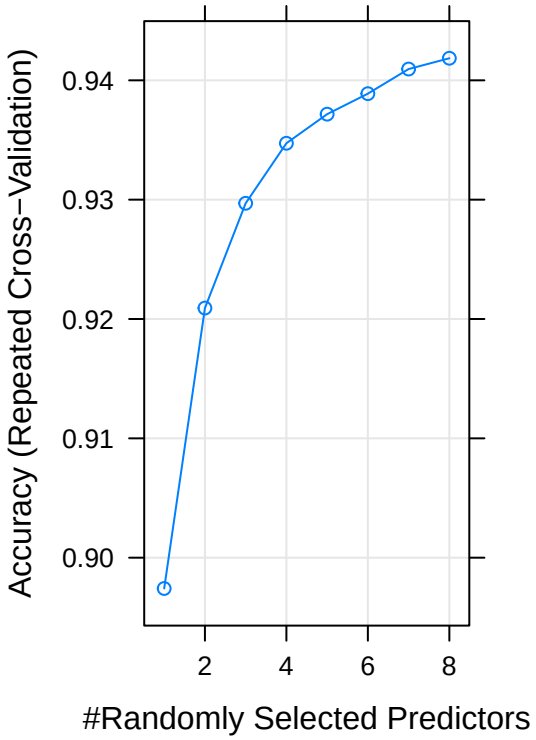

Figure 9

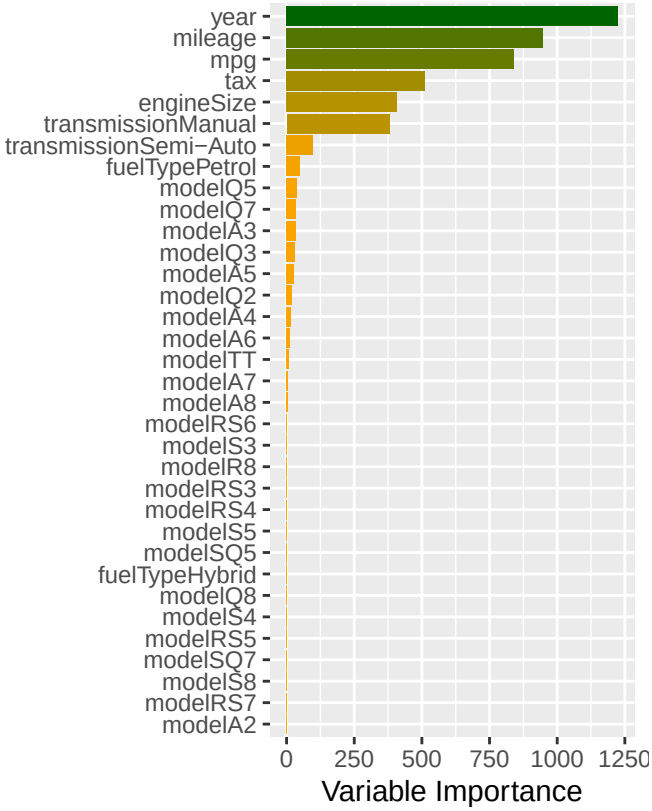

Figure 10

CONCLUSION

Findings

In this study, two models have been applied to forecast the price of Audi used cars. While we expect that some covariates correlate with the price, our Lasso and Random Forest models suggest that each variable influences the price but to a different extent. In particular, among the numerical variables, there is a strong positive association between the year of registration and car price, whereas mileage and miles per gallon are quite negatively associated with the price. Moreover, how much the price responds to a rise in the mileage is negatively impacted by when the car is registered. This means that a used car registered in a more recent year is likely to be more expensive, and the effect will be partially offset by an increase in the mileage. Lastly, our study also found that a hybrid car has a higher price or a manual transmission car is cheaper when controlling other categorical variables.

Related Work

Interestingly, what we have found is consistent with some of the ideas proposed in the study by Pudaruth, S. (2014) on the trend of vehicle production and consumer preferences. For example, while manufacturers tend to produce hybrid cars that depreciate slower than traditional ones, this is because the market becomes more aware of environmental concerns about the climate and the higher fuel efficiency of hybrid vehicles. In addition, year of registration, mileage, and mpg are considered to be key factors in determining the remaining value of a used car. The newer, less used and more fuel efficient the model, the higher the resale value.

Limitations

However, we would also like to admit three major areas of improvement in our study. The first limitation is that we derive our model based on limited data collected from an observational study indicating only association. We thereby raised one concern that it was yet not possible to conclude a cause-and-effect relationship between used car price and explanatory variables. After finding the correlation between the price and its covariates, we can work on further research to address this issue. Furthermore, we only have a small subset of factors that can affect the used car price in this study. One issue is that the inflation of car prices is not considered here when the models launched earlier could have a much lower initial price and resale value than the newer models could. Another prime importance is the rising fuel prices which are tied to the prices on fuel-efficient vehicles. Some other special factors which buyers attach importance to are the locality of previous owners, e.g., whether the car had been involved in serious accidents and whether it is a lady-driven car. The look and feel of the car could certainly contribute a lot to the price. As we can see, the price depends on a large number of attributes. Unfortunately, information about all these factors are not always available and the buyer must make the decision to purchase at a certain price based on a few factors only. Lastly, although small data sets can offer the advantage of sharp focus on particular issues, their narrow focus carries disadvantages as well. The challenges come from model building and the opportunity to synthesize the elements of regression learned one at a time from smaller data sets. With a larger and more richly structured data set, our future studies can dive deeper and further break down into more variables in order to better examine and explore if the rising price is the result of some other reasons.

REFERENCE

Kuiper, S. (2008). Introduction to Multiple Regression: How Much Is Your Car Worth?. Journal of Statistics Education, 16(3).

Pudaruth, S. (2014). Predicting the price of used cars using machine learning techniques. Int. J. Inf. Comput. Technol, 4(7), 753-764.

Puteri, C. K., & Safitri, L. N. (2020, February). Analysis of linear regression on used car sales in Indonesia. In Journal of Physics: Conference Series (Vol. 1469, No. 1, p. 012143). IOP Publishing.

Probst, P., & Boulesteix, A. L. (2017). To tune or not to tune the number of trees in random forest. J. Mach. Learn. Res., 18(1), 6673-6690.

Probst, P., Bischl, B., & Boulesteix, A. L. (2018). Tunability: Importance of hyperparameters of machine learning algorithms. arXiv preprint arXiv:1802.09596.