1. Regression Modelling

a. Build a regression model with the selected variables.

Solution: Attached in R script

b. Evaluate the regression model and carry out feature selection to build a better regression model. You need to try out at least 3 regression models to identify the optimal model.

Solution: I evaluated the regression model created in last step with parameters selected in Part B. I have used two evaluation methods R-Squared and Root Mean Squared Error (RMSE). I got R-Squared: 0.863162171173458 and Root Mean Squared Error (RMSE): 0.168578142286931 in the initial model without applying feature selection.

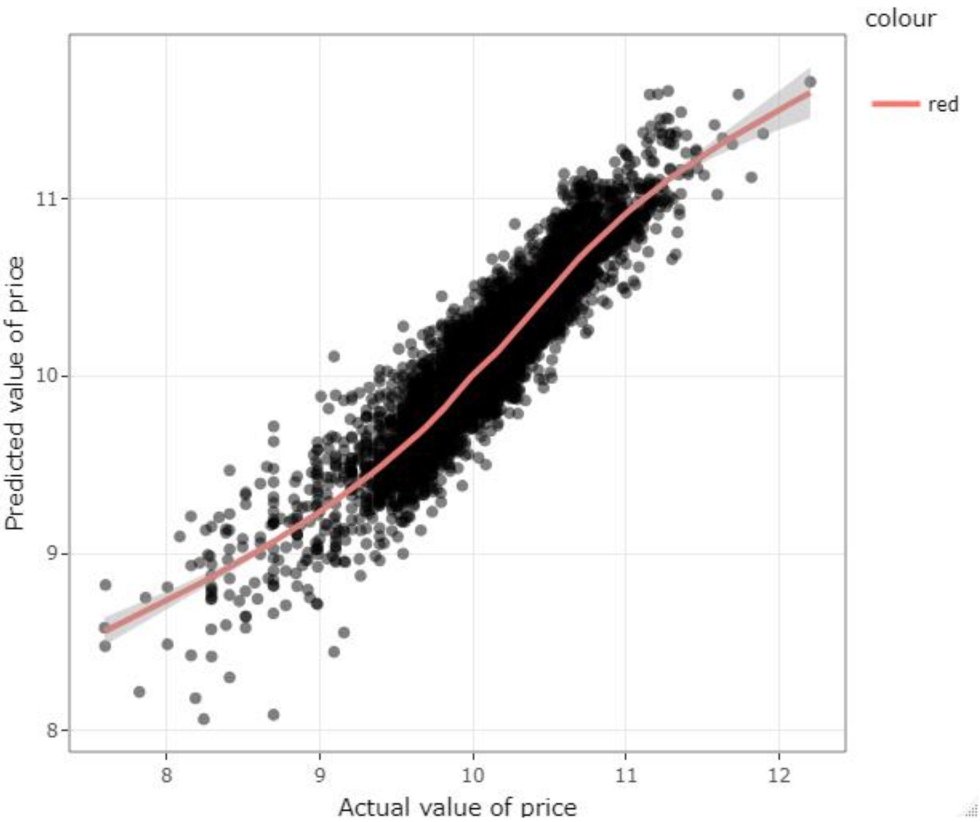

Figure 1 Predicted value of price v/s actual value of price; plot produced for initial regression model.

I carried out feature selection based on a plot produced by the *'Boruta Search'* by *'Boruta'* package. This package can show the importance of parameters for predicting target parameter in a hierarchical format.

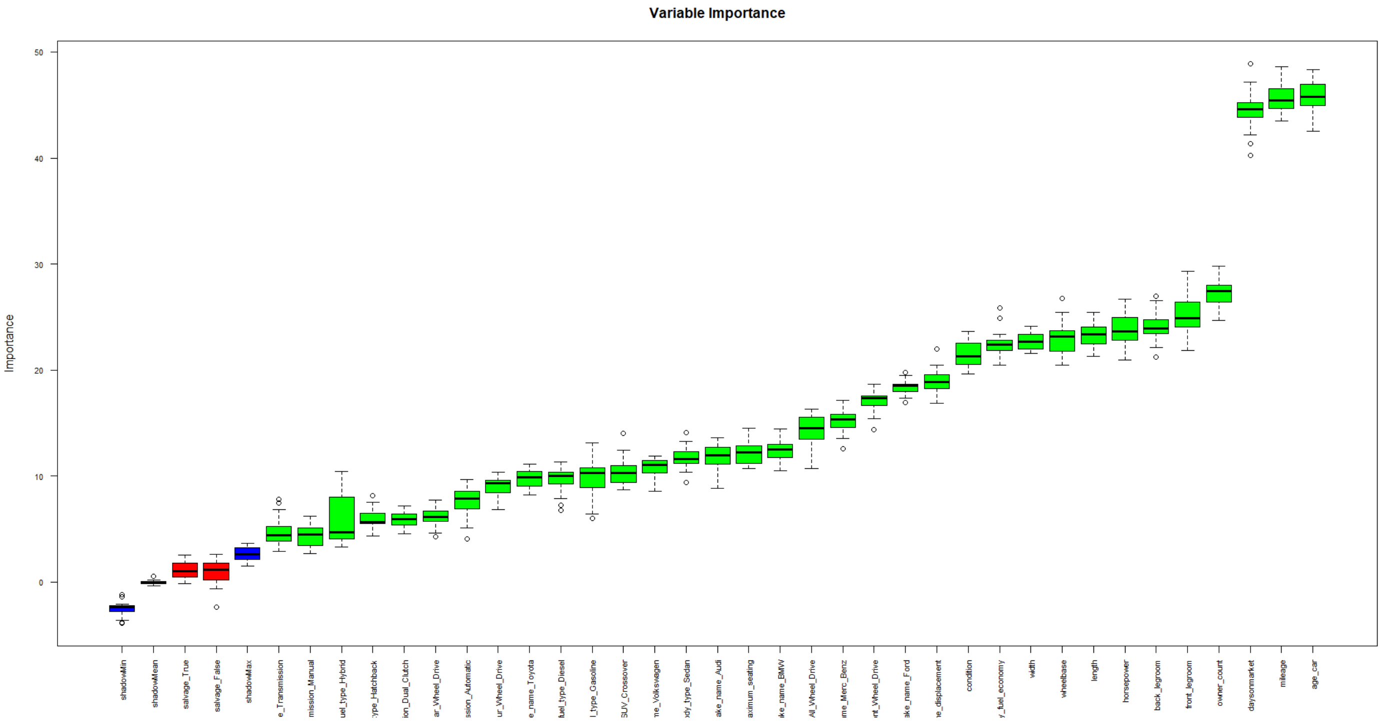

Figure 2 Plot showing the importance of parameters for predicting the target class (price).

<u>Model 1:</u>

In the first model based on feature selection, I selected the highly important features to include in our model *(price, age_car, mileage, daysonmarket, owner_count, front_legroom, back_legroom, length, horsepower, wheelbase, width, highway_fuel_economy, condition, engine_displacement).*

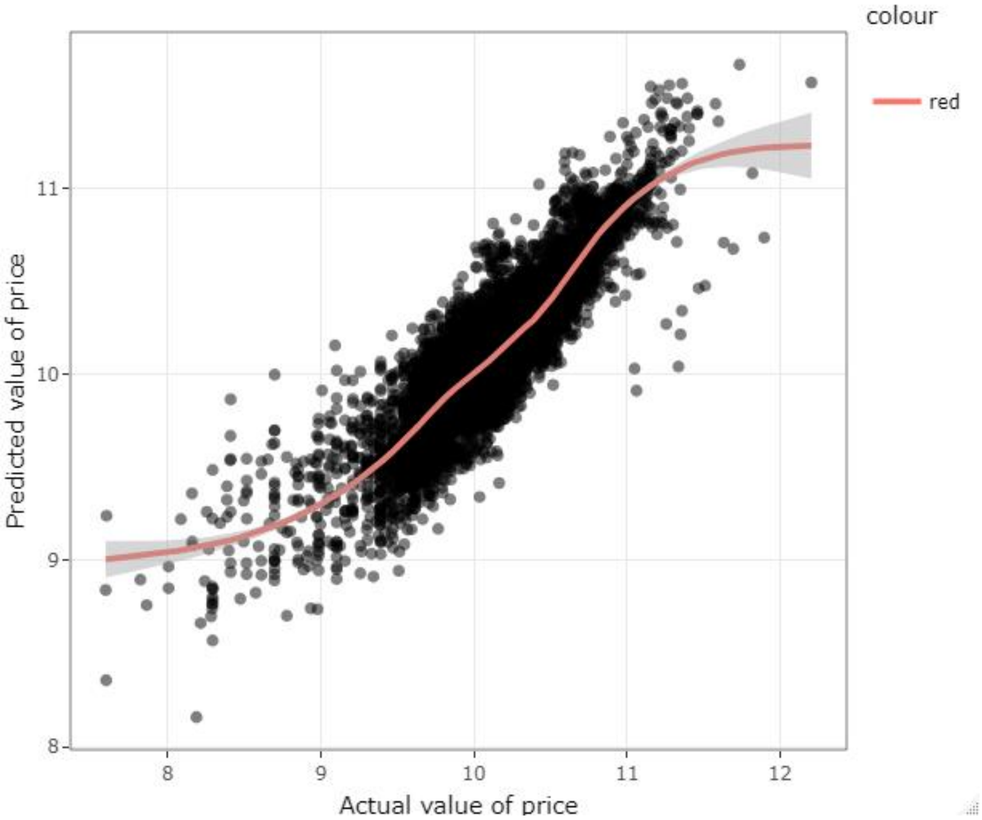

Figure 3 Plot for visual performance of Model 1 based on the feature selection of highly important parameters.

In the figure, we can see that the regression line is not perfectly diagonal. This indicates that there are some wrong predictions by the model. This is due to the model trained on the parameters which have outliers. A perfectly build regression model will have no outliers. But in *'used_cars'* data set we have a substantial number of outliers, and we can't drop them as they are a part of our study.

Model 2:

In this model, I added parameters *'make_name_Ford'* and *'wheel_system_Front_Wheel_Drive'*. This improved the prediction of the model quite significantly.

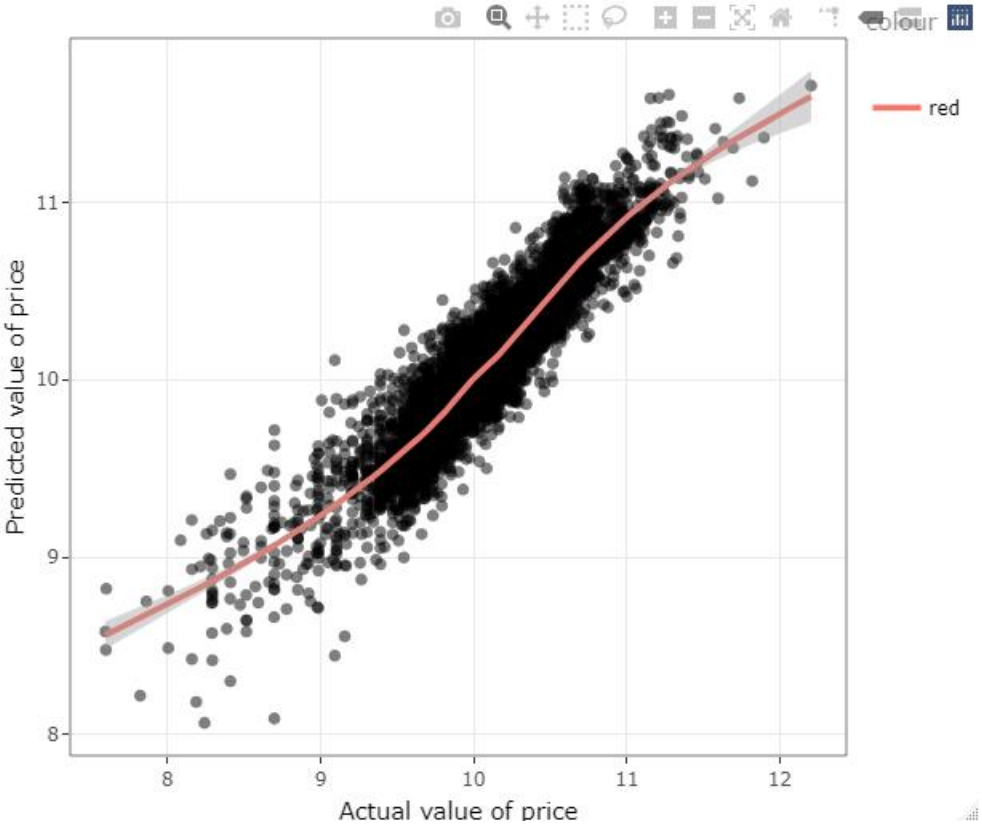

Figure 4 Plot for visual performance of Model 2; showing a better performance.

We can notice that the regression line is a bit straight as compared to the previous model which shows better prediction are made. This suggests that some features are highly essential in predicting the values of price. Adding even a few variables can make the model perform much better.

Model 3:

In this model, I used the parameters which are less important according to the plot in figure 2. I selected these variables for modelling; *(price, make_name_Ford, wheel_system_Front_Wheel_Drive, make_name_Merc_Benz, wheel_system_All_Wheel_Drive, make_name_BMW, maximum_seating, make_name_Audi, body_type_Sedan, make_name_Volkswagen, body_type_SUV_Crossover, fuel_type_Gasoline, fuel_type_Diesel, make_name_Toyota, wheel_system_Four_Wheel_Drive, transmission_Automatic, wheel_system_Rear_Wheel_Drive, transmission_Dual_Clutch, body_type_Hatchback, fuel_type_Hybrid, transmission_Manual, transmission_Continuously_Variable_Transmission)*.

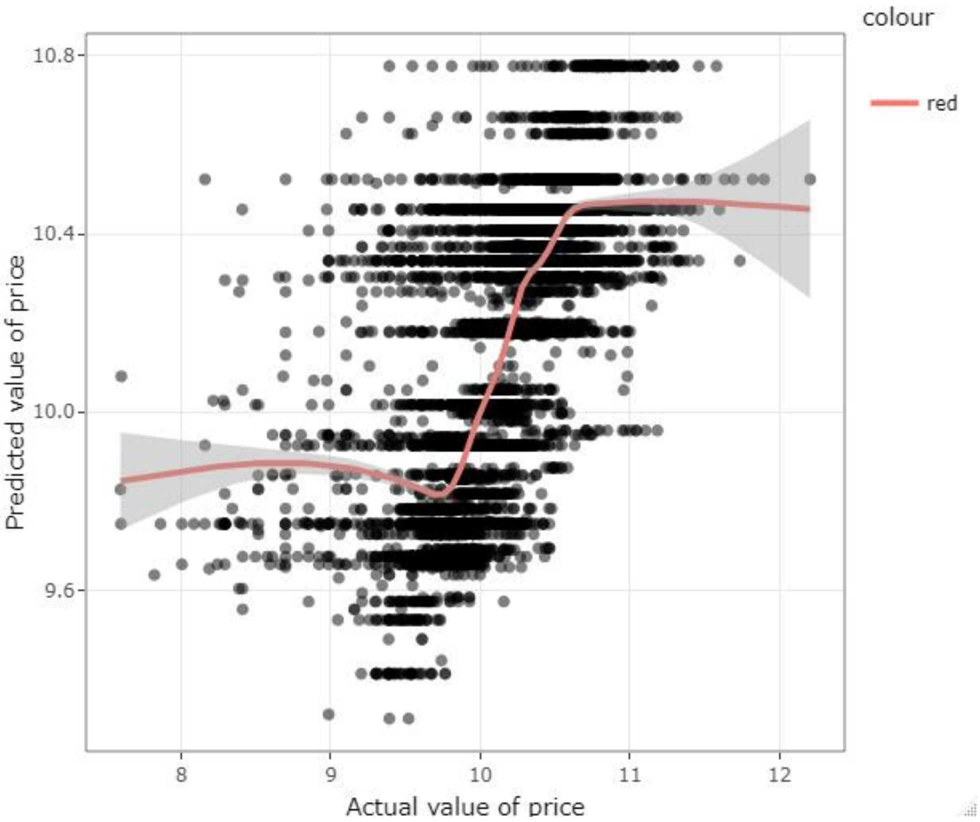

Figure 5 plot of model 3 is showing many wrong predictions.

As we can see, the number of variables used in this model are very high as compared to previous two models. However, after evaluating this model we can see that the model's performance is low.

For the three models we can observe that some parameters are highly important for predicting the target variable, while others are less useful in a prediction. Choosing the right parameters is an essential step for building an accurate prediction model.

c. Compare these regression models based on evaluation metrics and provide the formula for each regression model.

Solution:

Model	R Squared	Root Mean Square Error	Number of variables used
Model without Feature selection (F.S.)	0.863790648994337	0.168235027308904	37
Model 1 with (F.S.)	0.762231993275049	0.219216345285218	14
Model 2 with (F.S.)	0.841453802760523	0.181380066558472	16
Model 3 with (F.S.)	0.452116412771438	0.33545731563059	22

We can see from the evaluated results that the model without feature selection (F.S.) is not the best one as it uses 37 variables to get ~86 % accuracy. The model one with feature selection is giving less accuracy ~76% but has almost half the number of variables as compared to model without feature selection. Apparently, model two is the optimal model as it shows ~84 percent accuracy with 16 variables which is much better than both model one and model without F.S. The third model performs the worst ~45% accuracy with 22 variables used.

Formulas:

- Model without F.S.
  y ~ -10.92 + -0.13 * back_legroom + 0.15 * condition + -0.02 * daysonmarket + -0.24 * engine_displacement + 1.22 * front_legroom + -0.11 * highway_fuel_economy + 0.73 * horsepower + 2.17 * length + -0.24 * maximum_seating + -0.08 * mileage + -0.22 * owner_count + -0.89 * wheelbase + 2.21 * width + -0.33 * age_car + 0.22 * body_type_Hatchback + -0.1 * body_type_Sedan + NA * body_type_SUV_Crossover + 0.25 * fuel_type_Diesel + 0.02 * fuel_type_Gasoline + NA * fuel_type_Hybrid + -0.03 *

make_name_Audi + 0.02 * make_name_BMW + -0.53 * make_name_Ford + 0.09 * make_name_Merc_Benz + 0.08 * make_name_Toyota + NA * make_name_Volkswagen + -0.23 * salvage_True + NA * salvage_False + -0.02 * transmission_Automatic + 0.09 * transmission_Continuously_Variable_Transmission + 0.07 * transmission_Dual_Clutch + NA * transmission_Manual + 0.07 * wheel_system_All_Wheel_Drive + 0.03 * wheel_system_Four_Wheel_Drive + -0.04 * wheel_system_Front_Wheel_Drive + NA * wheel_system_Rear_Wheel_Drive

- Model 1 with F.S.
  y ~ 5.61 + -0.38 * age_car + -0.07 * mileage + -0.02 * daysonmarket + -0.26 * owner_count + -1.17 * front_legroom + -1.03 * back_legroom + -1.23 * length + 0.63 * horsepower + 3.73 * wheelbase + -0.28 * width + -0.14 * highway_fuel_economy + 0.16 * condition + 0.13 * engine_displacement
- Model 2 with F.S.
  y ~ -7.74 + -0.34 * age_car + -0.08 * mileage + -0.02 * daysonmarket + -0.24 * owner_count + 1.3 * front_legroom + -0.11 * back_legroom + -0.04 * length + 0.5 * horsepower + 0.44 * wheelbase + 2.27 * width + -0.03 * highway_fuel_economy + 0.16 * condition + 0.04 * engine_displacement + -0.54 * make_name_Ford + -0.11 * wheel_system_Front_Wheel_Drive

- Model 3 with F.S.
  y ~ 8.69 + -0.09 * make_name_Ford + -0.04 * wheel_system_Front_Wheel_Drive + 0.51 * make_name_Merc_Benz + 0.16 * wheel_system_All_Wheel_Drive + 0.39 * make_name_BMW + 0.75 * maximum_seating + 0.35 * make_name_Audi + -0.08 * body_type_Sedan + -0.07 * make_name_Volkswagen + -0.01 * body_type_SUV_Crossover + -0 * fuel_type_Gasoline + -0.26 * fuel_type_Diesel + NA * make_name_Toyota + -0.16 * wheel_system_Four_Wheel_Drive + -0.03 * transmission_Automatic + NA * wheel_system_Rear_Wheel_Drive + 0.09 * transmission_Dual_Clutch + NA * body_type_Hatchback + NA * fuel_type_Hybrid + -0.13 * transmission_Manual + NA * transmission_Continuously_Variable_Transmission

2. Decision Tree Modelling

a. Build a decision tree with the selected variables.

Solution: Attached in R script.

b. Evaluate the decision tree model and carry out pruning to build a better decision tree model. You need to try out at least 3 decision trees to obtain the optimal tree.

Solution: I evaluated the decision tree model (Model one) created in last step considering same parameters that were used in regression modelling. I used Root Mean Square Error (RMSE) to evaluate the accuracy of the models. I got RMSE: 0.217622527686824.

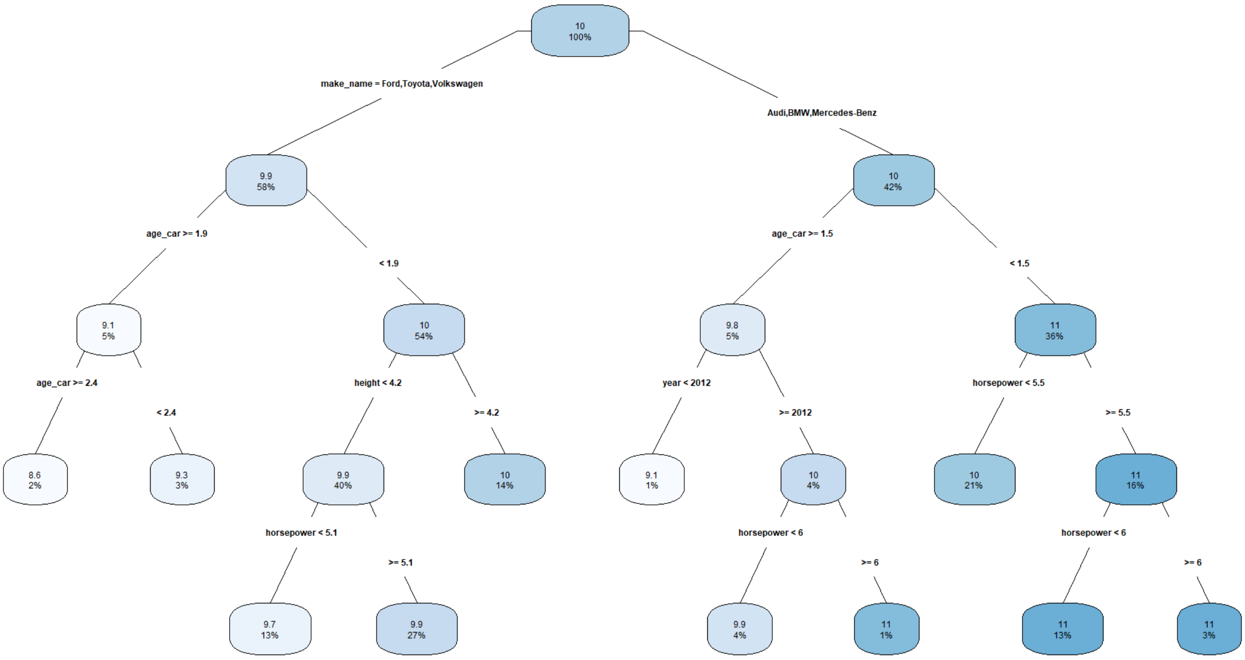

Figure 6 Decision tree before pruning.

After pruning the decision tree with complexity parameter (C.P.) value 0.01, model one shows a lower RMSE: 0.00198577010683942. In model two after applying pruning with CP value 0.02, model two decision tree showed RMSE: 0.00425507567859576. In model three with CP value 0.04, the decision tree showed RMSE: 0.00304833221742191.

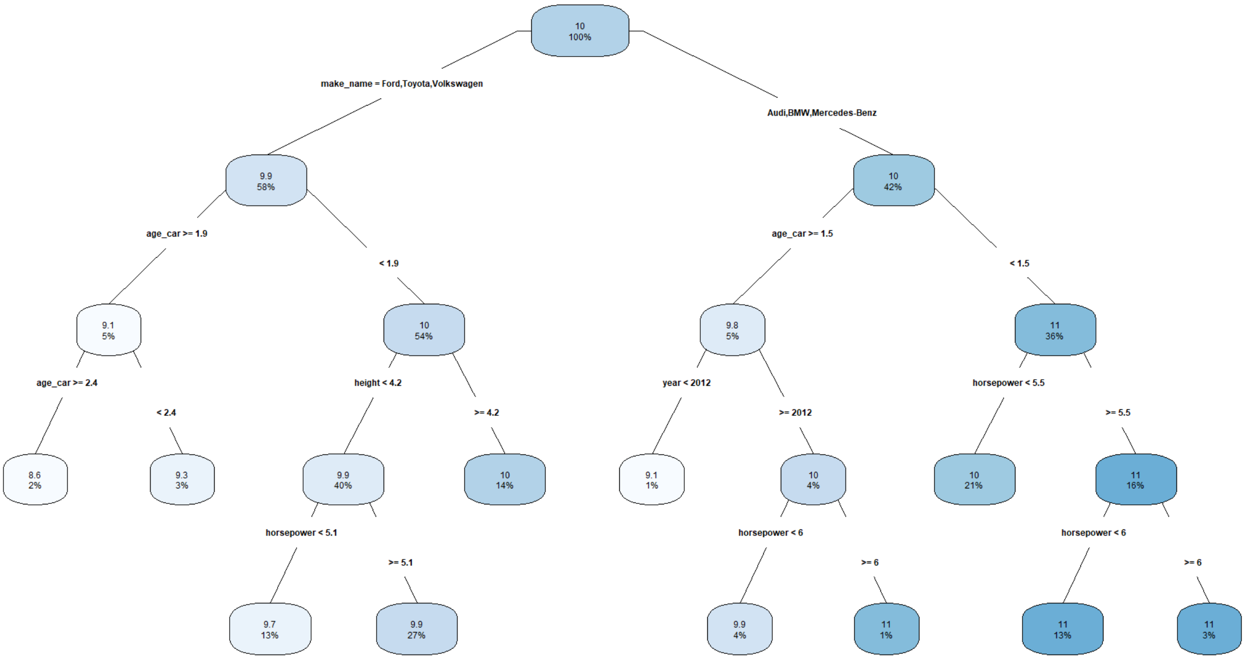

Figure 7 Decision tree after pruning with 0.01 CP value.

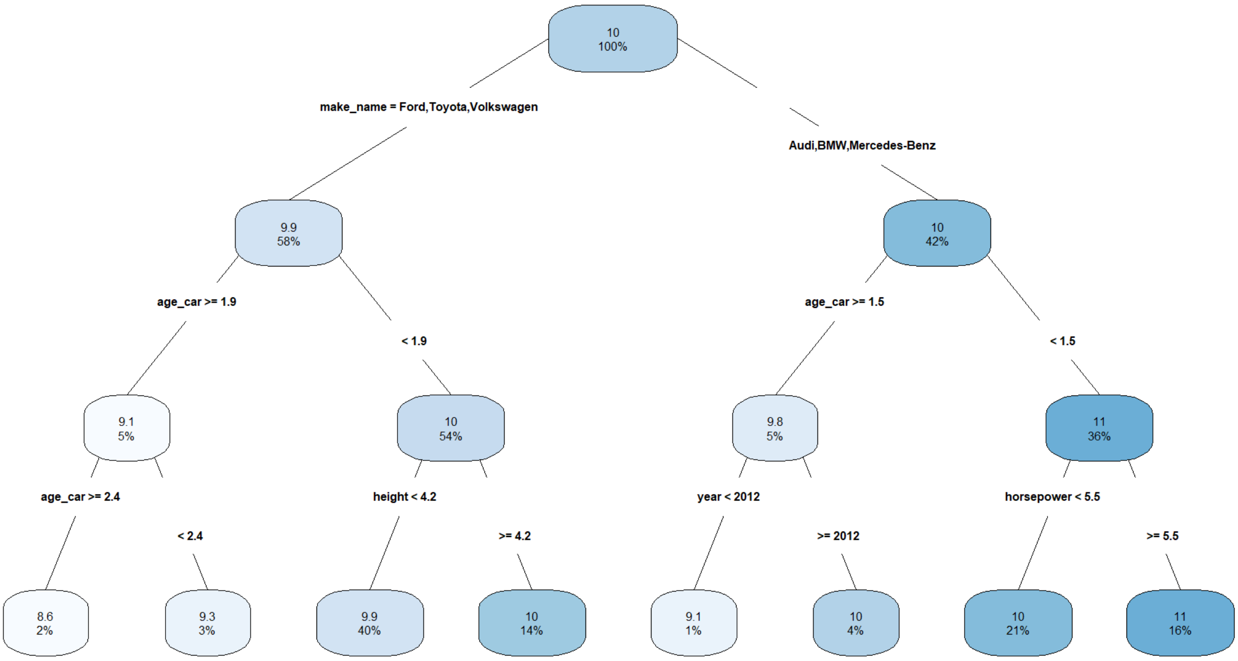

Figure 8 Decision tree after pruning with 0.02 CP value.

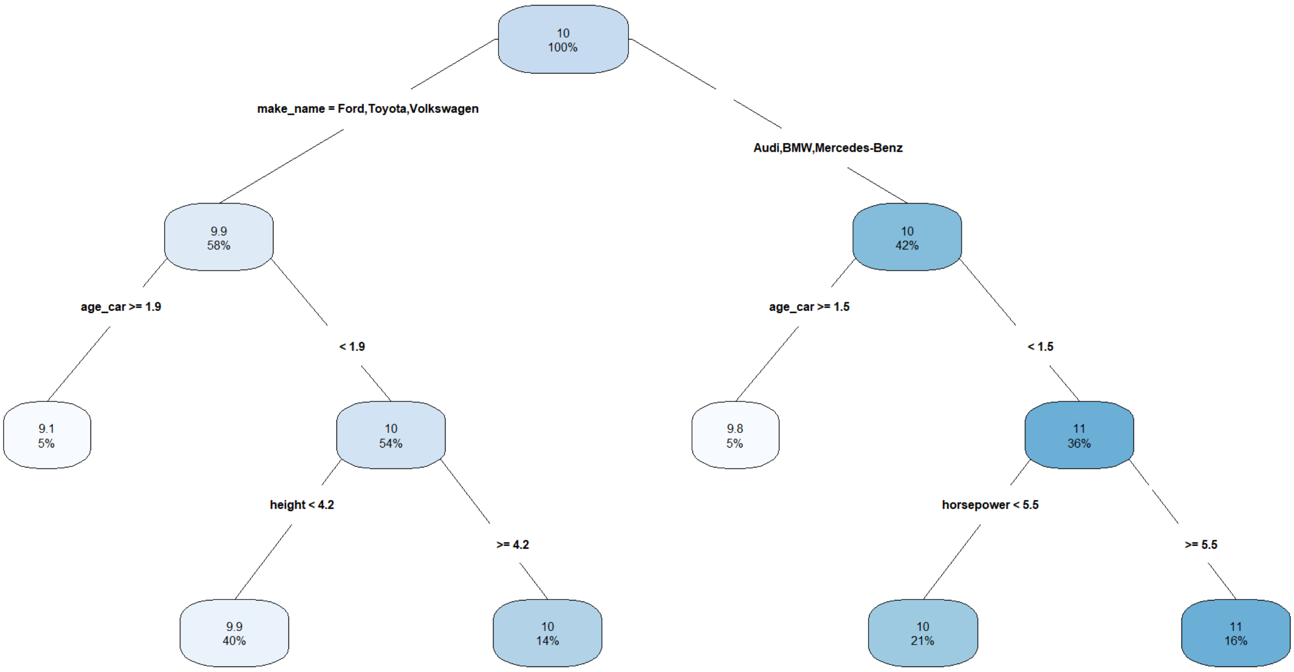

Figure 9 Decision tree after pruning with 0.04 CP value.

c. Compare these decision tree models based on evaluation metrics and provide the tree plot for each model and explain the outputs.

Solution: We can see from the plots and the RMSE value that all the pruned decision trees have much better accuracy that the decision tree without pruning.

Model	RMSE value	Number of nodes	Number of levels (Depth)
Model without pruning	0.217622527686824	21	5
Model one with pruning (CP =0.01)	0.00198577010683942	21	5
Model two with pruning (CP =0.02)	0.00425507567859576	15	4
Model three with pruning (CP =0.04)	0.00304833221742191	11	4

However, we can observe that the Model with CP value 0.01 has the lowest RMSE value. This means that decision tree with CP value 0.01 is the optimal decision tree. Also, this decision tree is deeper or has more nodes as compared to rest of the two models with a higher CP value.

3. Model Comparison

a. Why do we need to build several models in both regression and decision trees (as requested in question 1 and 2)?

Solution: We need to build many models in both regression and decision trees because models with different parameters perform differently. We need to check the accuracy of different models to decide which model is the optimal using evaluation methods. It is also to avoid overfitting and underfitting as using parameters or data points with less or more variability in training phase of model can result in model trained on a narrow or a huge scale. Adding too much or limited information can cause underfitting or overfitting as model will not be able to work as per the business problem. So, we perform many steps like transformations, replacing missing values, removing outliers to make sure that the data is best fit for our regression and decision tree models. We used pruning method for decision trees to get the optimal tree.

b. Compare the accuracy of the selected (optimal) regression model and (optimal) decision tree and discuss and justify the most suitable predictive model for the business case.

Solution: We used R-square in regression modelling and RMSE method in both regression and decision tree modelling to evaluate the results of the models. As RMSE is the common evaluation method we can consider it for comparing the accuracies between the optimal regression and decision tree models.

Model	RMSE value
Regression model	0.181380066558472
Decision tree model	0.00198577010683942

We can clearly see that the decision tree performed almost 100 times better than the regression model. For the business case we can consider the decision tree as the most suitable predictive model as it outperforms all other methods and parameter settings we have tried in our study. The reason for such a big difference in the performance is that decision tree can create much more complex boundaries to decide the target variable. In this article it is clearly discussed how decision trees are generally better for prediction tasks due to decision boundaries.

4. Modelling and Comparison with a Filtered Data Set

a. Filter and select rows with only one specific make_name (i.e., BMW) from the data set with selected variables.

Solution: Attached in R script.

b. Redo the question 1 and 2 based on the filtered data (only need to build one model each based on the best model setting from Q1 and Q2) (Note: if make_name is included your previously selected variables, remove it before build any model, since now make_name only contains one value, and will not contribute to prediction)

Solution: Attached in R script.

c. Compare the performance of regression model before filtering and after filtering, and the decision tree model before filtering and after filtering. What do you find? Discuss your findings

Solution:

Model	**RMSE Before Filtering**	**RMSE After Filtering**	**R squared before filtering**	**R squared after filtering**
Regression Model	0.181380066558472	0.125677887284789	0.841453802760523	0.911638637213219
Decision Tree	0.00198577010683942	0.00114507693232625	-	-

As we can see in the r script that after applying filter for the make model BMW, the number of data points reduced from 23243 to 4160 rows. This reduced the size of the data set. So, the model has more similar and small data set to build on and will predict better for cars of BMW brand. Therefore, we see that the RMSE reduced in both regression model and decision tree after filtering. R squared value also increased to ~91 after filtering. This shows a higher accuracy in both models. However, now the model is only built to predict values for BMW brand cars and if we apply it to predict used car prices for other brands it will not give good results. This is called overfitting. So, generally we avoid overfitting and consider a wider range of parameters for training the model for a more generalized data prediction.

References

1. Contributor, G. (2019, December 16). *Why decision trees are more flexible than linear models, explains Stephen Klosterman*. Packt Hub. https://hub.packtpub.com/why-decision-trees-are-more-flexible-than-linear-models-explains-stephen-klosterman/