Predictive Analytics Project

Group 3: Nguyễn Huỳnh Đức, Hoàng Lê Mỹ Anh, Lại Thị Hạnh, Đào Lê Huy, Lê Bá Huy

1. Clean data:

After preprocessing data in R, we have 3,826 observations left in training data set and 1,014 observations left in testing data set.

clean_training_datas
et.csv

clean_testing_dataset
.csv

2. Build A linear Regression Model:

a. After running full linear regression model, we have MSE for training set and testing set as below:

$MSE_{training}$= 64,652,528,716 $MSE_{testing}$= 130,608,343,787

Because $MSE_{testing}$ is much higher than $MSE_{training,}$ it suggests us about overfitting in the model.

b. From results in R, we have p-value of F-test for the validity of the model is $2.2*10^{-16}$ which is much smaller than threshold of 5%. As results, we can reject H0, the null hypothesis which means there is no predictive relationship between the X variables and Y in the population.

c. It is noticeable that over 13 explanatory variables, two variables: fuel and seats should be excluded from the model. Because p-value for T-test for seats variable is 0.334031 much higher than threshold of 5%, it suggest accepting the null hypothesis that the coefficient of seats variable is equal to zero. In other words, there is no relationship between the seats variable and the selling_price variable.

It is quite the same with fuel variable which has categorical data, then it has three dummies: fuelDiesel, fuelLPG, fuelPetrol. However, the p-value for T-test of these whole three variables are much larger than threshold of 5% (with p-value = 0.839458, 0.087947, 0.198724 respectively). Therefore, we can conclude that there is no relationship between the fuel variable and the selling_price variable.

We respectively exclude each of these two variables (fuel, seats) to see whether the remaining variable can be significant or not. However, these two variables are still insignificant in the two models. Therefore, it is reasonable to exclude both fuel and seats variables.

d. After excluding fuel and seats variables, we have final with 11 explanatory variables left.

e. In our final model, we have tables of significant variables and its coefficients as below:

Variable	Coefficient
Intercept	-8.412e+07
nameAudi	5.052e+05
nameBMW	1.138e+06
nameChevrolet	-4.836e+05
nameDatsun	-4.826e+05
nameFiat	-4.910e+05
nameForce	-4.571e+05
nameFord	-3.745e+05
nameHonda	-3.611e+05
nameHyundai	-3.708e+05
nameJaguar	6.583e+05
nameLand	1.020e+06
nameLexus	3.383e+06
nameMahindra	-4.491e+05
nameMaruti	-3.134e+05
nameMercedes-Benz	7.044e+05
nameMitsubishi	-3.215e+05
nameNissan	-4.037e+05
nameRenault	-4.190e+05

nameSkoda	-3.842e+05
nameTata	-5.091e+05
nameVolkswagen	-5.093e+05
year	4.216e+04
km_driven	-1.005e+00
seller_typeIndividual	-6.612e+04
transmissionManual	-8.862e+04
ownerFourth & Above Owner	-7.195e+04
ownerSecond Owner	-5.155e+04
ownerTest Drive Car	2.570e+06
mileage	-8.128e+03
engine	5.527e+01
max_power	3.512e+03
Nm	1.114e+03
rpm	-3.291e+01

f,

- Intercept Coefficient: The intercept coefficient is -8.412e+07. It represents the expected value of the target feature when all predictor variables are set to zero. However, this intercept is negative, it is not meaningful for selling price of car.

- Quantitative Feature Coefficient: The coefficient for Nm is 1.114e+03. This suggests that for every one-Nm increase in Nm, the car selling price is expected to increase by 1.114e+03.

- Qualitative Feature Coefficient: For the qualitative features "owner", there are 5 categories, so this variable has 4 dummies. The "First Owner" is reference category. The coefficient of the dummies ownerFourth & Above Owner, ownerSecond Owner, ownerTest Drive Car and ownerThird Owner are -7.195e+04, -5.155e+04, 2.570e+06 and -3.689e+04, respectively which means these are the differences between each dummy with the reference category.

3. Build An Elastic Net Model:

a. Running alpha from 0 to 1 to check for the lambda min in each case and then choose the best alpha. As a results, we can have alpha = 0.5, lambda = 90.58914

b. We have MSE for training set and testing set as below:

$MSE_{training}$= 64,811,887,903 $MSE_{testing}$= 130,182,362,516

Because $MSE_{testing}$ is much is much higher than $MSE_{training}$, it suggests us about overfitting in the model.

c. In our model, we have tables of variables and its coefficients as below:

Variable	Coefficient
(Intercept)	-8.466002e+07
nameAshok	-6.825559e+05
nameBMW	7.315198e+05
nameChevrolet	-8.347847e+05
nameDaewoo	-2.762255e+05
nameDatsun	-8.409786e+05
nameFiat	-8.498878e+05
nameForce	-8.277192e+05
nameFord	-7.302347e+05
nameHonda	-7.214740e+05
nameHyundai	-7.325433e+05
nameIsuzu	-2.495416e+05
nameJaguar	2.525342e+05
nameJeep	-1.75020e+05
nameKia	-4.485604e+05
nameLand	6.407468e+05
nameLexus	2.987380e+05
nameMahindra	-8.035935e+05
nameMaruti	-6.771883e+05
nameMercedes-Benz	2.986663e+05
nameMG	-2.021403e+05
nameMitsubishi	-6.962181e+05
nameNissan	-7.649080e+05
nameOpel	-4.160889e+05
nameRenault	-7.777717e+05

nameSkoda	-7.528326e+05
nameTata	-8.648588e+05
nameToyota	-4.358631e+05
nameVolkswagen	-8.759665e+05
nameVolvo	-1.634358e+05
year	4.259903e+04
km_driven	-9.569541e-01
fuelDiesel	-1.727186e+04
fuelLPG	1.396170e+05
fuelPetrol	5.708926e+04
seller_typeIndividual	-6.598125e+04
seller_typeTrustmark Dealer	-6.399539e+04
transmissionManual	-9.169778e+04
ownerFourth & Above Owner	-6.806257e+04
ownerSecond Owner	-5.011748e+04
ownerTest Drive Car	2.613398e+06
ownerThird Owner	-3.626063e+04
mileage	-6.111051e+03
engine	7.338783e+01
max_power	3.394097e+03
Nm	1.328410e+03
rpm	-5.309490e+01
seats	-6.875749e+03

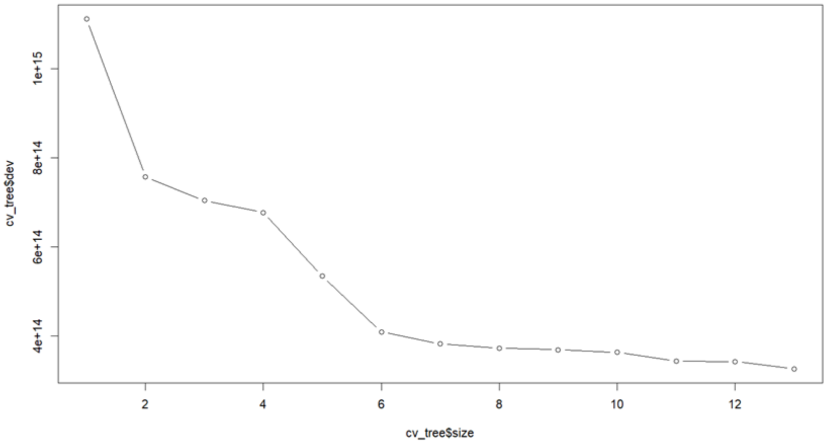
4. Build A Regression Tree Model:

a. After running regression tree model, we have MSE for training set and testing set as below:
$MSE_{training}$= 51,297,653,199 $MSE_{testing}$= 126,027,267,454

Because $MSE_{testing}$ is twice as high as $MSE_{training}$, it suggests us about overfitting in the model.

b. The depth of tree = 5, and the features involved in the splits: "max_power", "year", "Nm", "name", "km_driven", "mileage".

c. Based on cross-validation (5 folds), check for the best size of the tree and prune it accordingly.



As we can see, even though the tree size of 13 has the minimum cross-validation error, the amount of deviation decrease compared to the tree size of 6 is immaterial. So, we will prune the tree from our original tree to get the optimal tree size of 6.

d. After running pruned tree model, we have MSE for training set and testing set as below:
$MSE_{training}$= 81,263,513,600 $MSE_{testing}$= 137,464,205,071

Because $MSE_{testing}$ > $MSE_{training}$, the pruned tree model still overfitted but less than original tree models.

The depth of tree = 3, and the features involved in the splits: "max_power", "year", "km_driven".

e. Since the improvement in interpretability outweighs the increase in MSE, the pruned tree delivers better results than the original tree, so pruning is advised.

5. Build A Random Forest Model:

a. For the intial random forest model with mtry =4 and ntree = 10, we have the following results:

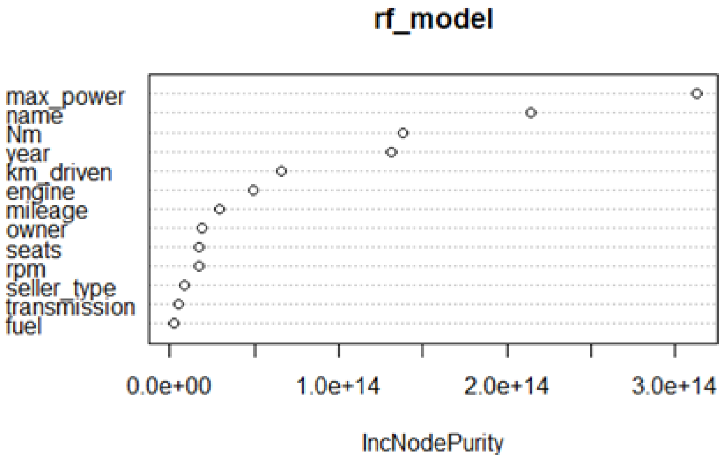
$\text{MSE}_{\text{training}}$ = 12,189,401,735 $\text{MSE}_{\text{testing}}$ = 50,729,320,651

As can be seen, the MSE of the random forest model on the test datset is significantly higher than the MSE of the training set, suggesting that our model is overfitted. This is expected given the small number of trees and the number of selected variables for each tree.

b. For random forest model with mtry = 4 and ntree = 10, we have:

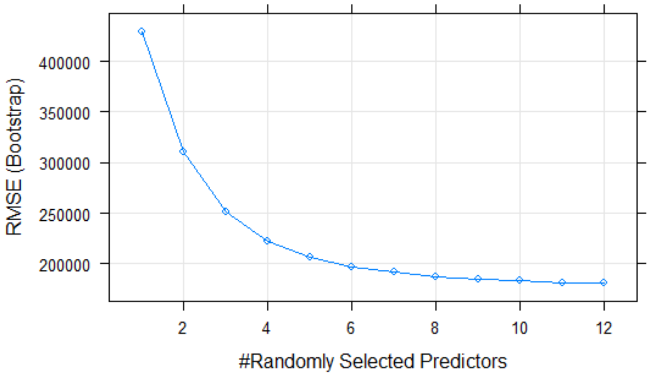
Mean of squared residuals = 45,297,267,617 % Var explained = 84.41

As we can see from the results of the model, the mean of squared residuals is considerably high given that the mean value of response variable is 553,065.54. This implies that model fit is not robust. Nevertheless, the percentage of variance explained is 84.41%, indicating that the variance in response variable can be explained quite well by the model.



Based on the plot, four variables “max_power”, “name”, “Nm” and “year” make the largest contribution to the model. This result suggests that the selling price of a car is largely determined by the car make, engine power and vehicle age. It aligns with the domain knowledge in car industry since these factors represent the car quality and overall performance.

c. Using 5-fold cross validation, we find that mtry = 12 produces the least RMSE, so we can choose 12 as the mtry for our optimized our model. We continue to tune hyperparameter ntree and observed that the best ntree = 250, with mean value of RMSE = 168,647.5.

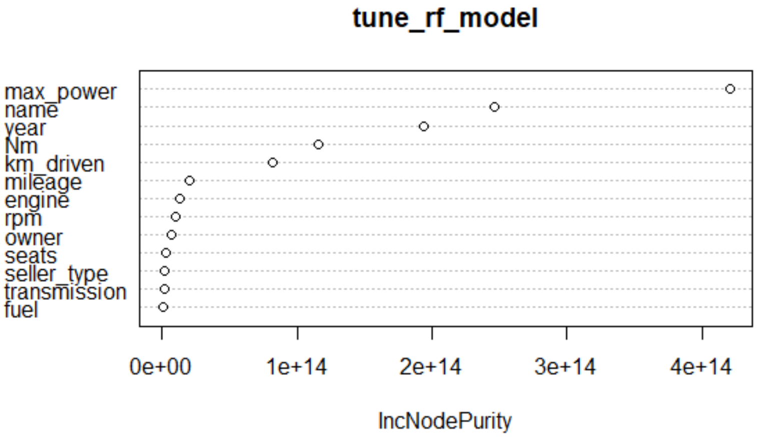


d. From the optimized random forest model with mtry = 12 and ntree = 250, we obtained:

Mean of squared residuals = 29,124,338,206 % Var explained = 89.98

MSE$_{\text{training}}$ = 5,436,742,026 MSE$_{\text{testing}}$ = 52,362,303,980

This result implies that the optimized random forest model still suffers from overfitting.



In addition, variables “max_power”, “name”, “Nm” and “year” continue to make the highest contribution to the explanatory ability of the model. This is consistent with our finding and comment in part 5b.

e. After tuning the hyperparameter mtry and ntree, we observe that the tuned random forest model suffers from overfitting even heavier than the original random forest model, while the percentage of variance in response variable explained by the model only increases slightly. Taking this trade-off into consideration, we can conclude that the tuned model does not deliver better results than the original model.

6. Best Model Selection

Based on 03 criteria – Robustness, Performance and Interpretability, we consider the regression tree model with mtry = 4 and ntree =10 as the best since all our models suffer from overfitting; however, the MSE on both the training and testing datasets of the random forest model are the lowest. This implies that the model provides the best predictive performance among all models. Furthermore, the small number of trees make it easy for interpretation.