Before fitting the data into the logistic regression model to predict the car price ranges, we took following steps:

* We manipulated the data by excluding data of top-luxury car brands, including Ferrari, Rolls-Royce, Maybach, McLaren, Lamborghini, Aston Martin, Jaguar, Bentley, Porsche, Maserati. Prices of cars of these brands are always on the top, classifying such luxury cars would make little sense.
* To create a binary outcome of the price, we determined the threshold of the price to divide prices into two ranges: “affordable” and “expensive” by taking the median of prices after excluding top-luxury car brands.
* We removed columns with irrelevant features which either are not helpful predictors or contain large amount of missing values.
* We converted all categorical variables into factor type.
* We checked the assumptions of logistic regression, including linearity between independent variables and the logit of outcome and multicollinearity between independent variables. For linearity between independent variables and logit of outcome, we checked by scatterplot. Linearity relationship between each independent variable and logit of outcome was confirmed. For multicollinearity, we checked linear relationship among 3 variables: Displacement\_cm3, Power\_HP and Mileage\_km by scatterplot and confirmed by VIF values over 2.5, which is considered worth concerning (Johnston et al., 2018). Multicollinearity was found between Displacement\_cm3 and Power\_HP variables, we chose to remove Displacement\_cm3. (See appendix A for two assumptions checking)
* We used visual inspection to select the most important categorical variables for the logistic regression: Production\_year, Vehicle\_brand, Type and Doors\_number.
* We subsetted the data for logistic regression using the categorical predictors including Production\_year, Vehicle\_brand, Type, Doors\_number, and continuous predictors which do not violate multicollinearity assumption: Power\_HP and Mileage\_km.

After choosing the most important variables for logistic regression model, we proceeded to fit the data into logistic regression model.

* We splitted the data into training and test set with 80% of data for training set and 20% for test set.
* For logistic regression model, we have 6 predictors and 1 binary outcome. We fitted logistic regression model into the training set, predicting whether the price\_label (price range) of a car is “affordable” or “expensive” or not based on all other variables: production year, vehicle brand, type, doors number, horse power and mileages.
* We trained the logistic regression model with reference level of the price (price\_label) set to “affordable” and used 5 fold cross validation method. For each independent variable, we set following reference levels for easier interpretation: vehicle brand’s reference level is “Daewoo”, type of car’s is “compact”, door number’s is 2, production year’s is 2021.
* We predicted the price outcomes on the test set.
* We checked the performance of the logistic regression model in predicting the price range using ROC curve.

Interpretation of our main findings for logistic regression model:

As reference level of price is set to “affordable” and production year to 2021, as for cars with production years between 1983 and 1984 and between 1999 and 2020, they are less likely to be expensive compared to cars produced in 2021.

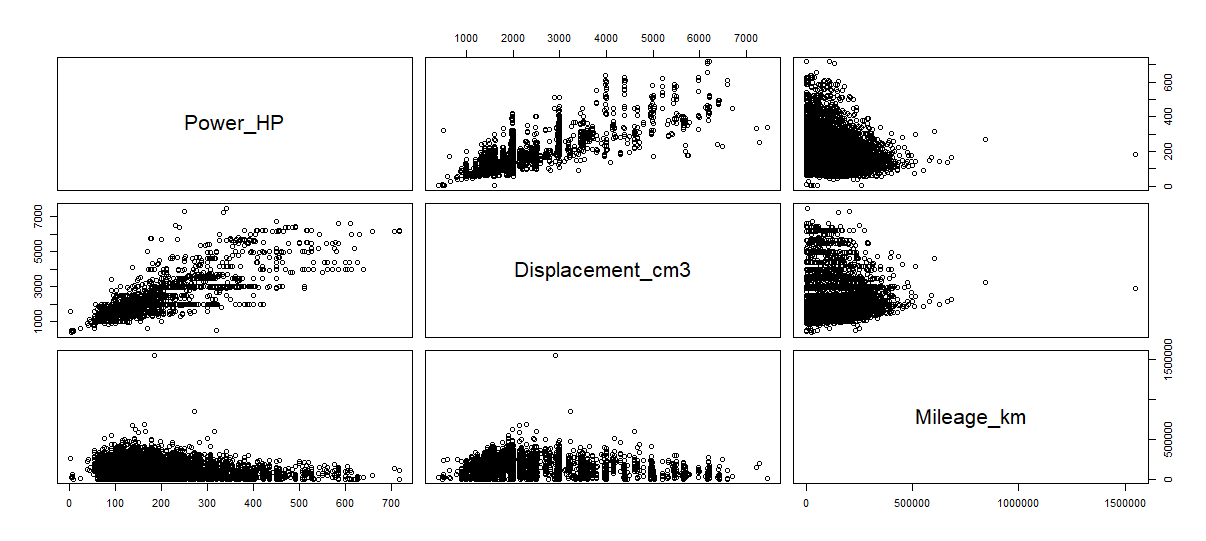
As for city cars and small cars, compared to compact cars, they are less likely to be expensive. Convertibles, coupes, minivans, sedans, station wagons and SUV are more likely to be expensive compared to compact cars.

For one unit increase in mileages driven by a car, the log odds of a car being expensive decreases by 0.00001687 For one unit increase in horse power, the log odds of a car being expensive increases by 0.02313.

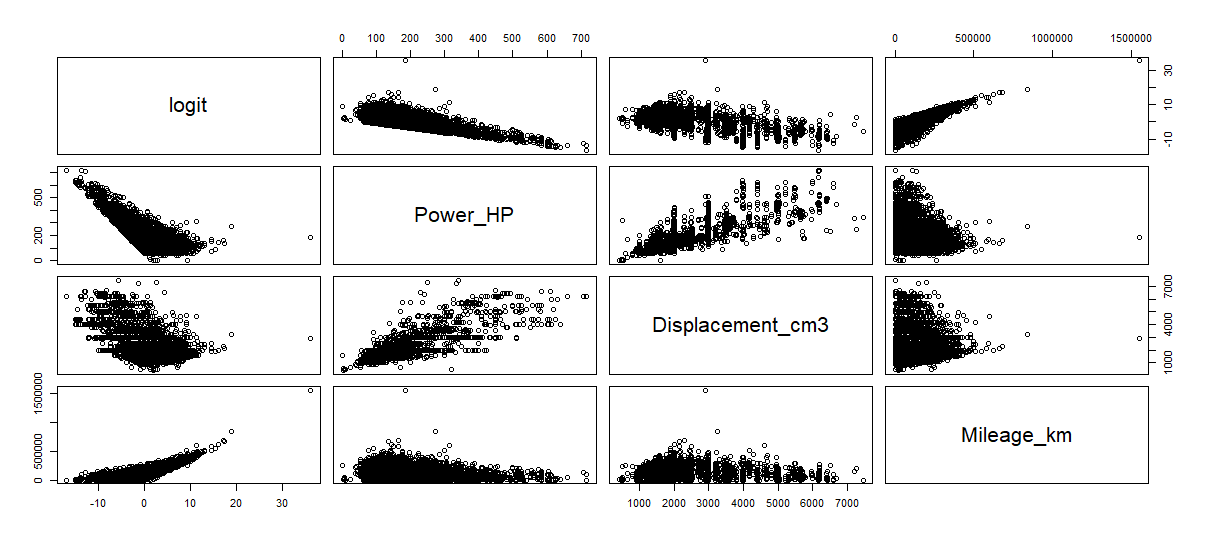
**Conclusion**

Overall, our model has proved the accuracy score of 92.98%, recall score (sensitivity) of 92.84%, referring to the ability of our model to correctly identify cars with affordable price, and precision score (specificity) of 93.13%, referring to the ability of our model to correctly identify expensive cars. Our model showed that production year, type of cars, mileages in kilometers driven, manual transmission, hybrid and diesel fuel type and horse power are useful predictors, while doors number and vehicle brands are not. (see appendix B for our confusion matrix and table of coefficients).

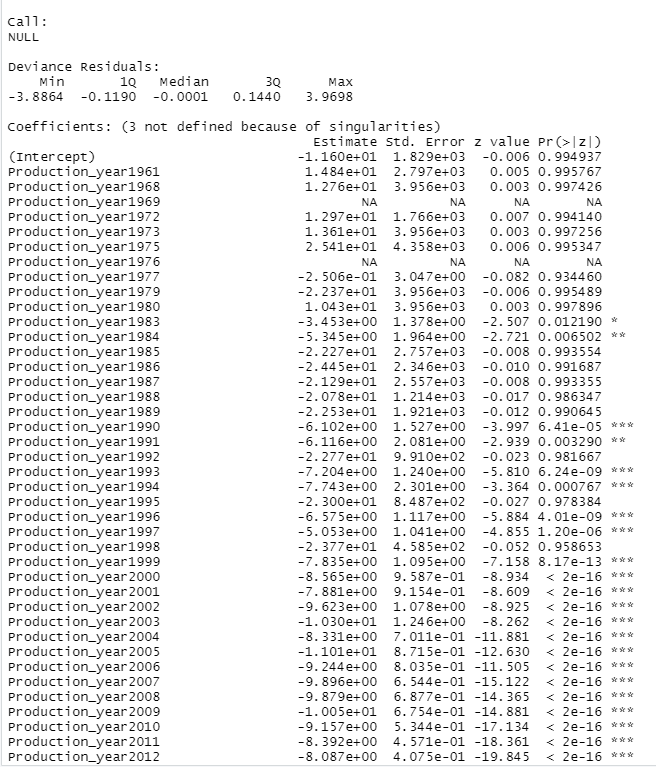
Appendix A

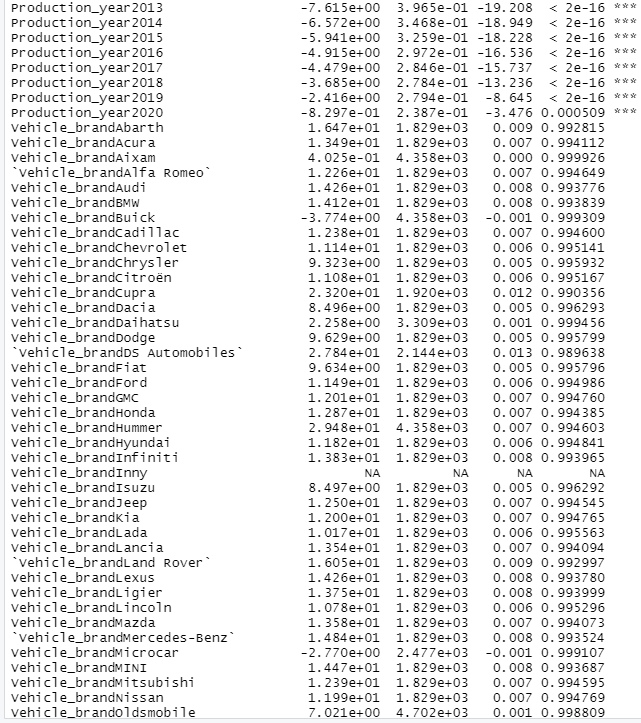
 

Multicollinearity checking

Checking of linearity between independent variables of logit of outcome

Appendix B





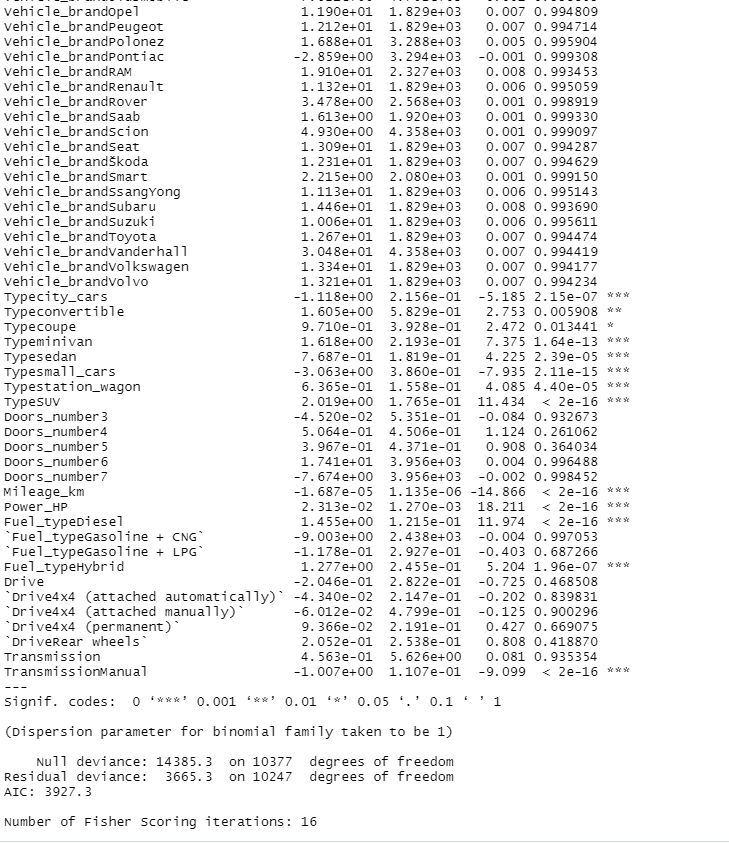
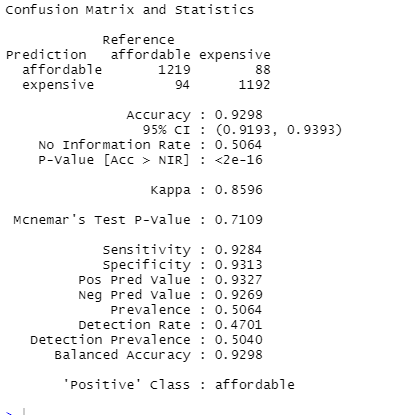


Table of coefficients



Confusion matrix

**References**:

Johnston R, Jones K, Manley D. Confounding and collinearity in regression analysis: a cautionary tale and an alternative procedure, illustrated by studies of British voting behaviour. Qual Quant. 2018;52(4):1957-1976. doi:10.1007/s11135-017-0584-6

library(broom)

library(caret)

library(dbplyr)

library(tidyverse)

library(plotmo)

library(glmnet)

library(devtools)