Data Science in R: Analysis of Vehicles

William Wager Johnsen

26/12/2019

Objective

- Identifying whether a vehicle has relatively environmental good milage.
- Identify which characteristics that has a statistical significance on the mileage per gallon.

Description of data

- Dataset: Auto-Mpg Data.
- Origin: This dataset was taken from the StatLib library, which is maintained at Carnegie Mellon University. The dataset was used in the 1983 American Statistical Association Exposition.
- Date: July 7, 1993.
- Date of observations: 1970 to 1982.
- Assumption: A vehicle has a good consumption if it can do 23 miles per gallon of gasoline.

Process of Analysis

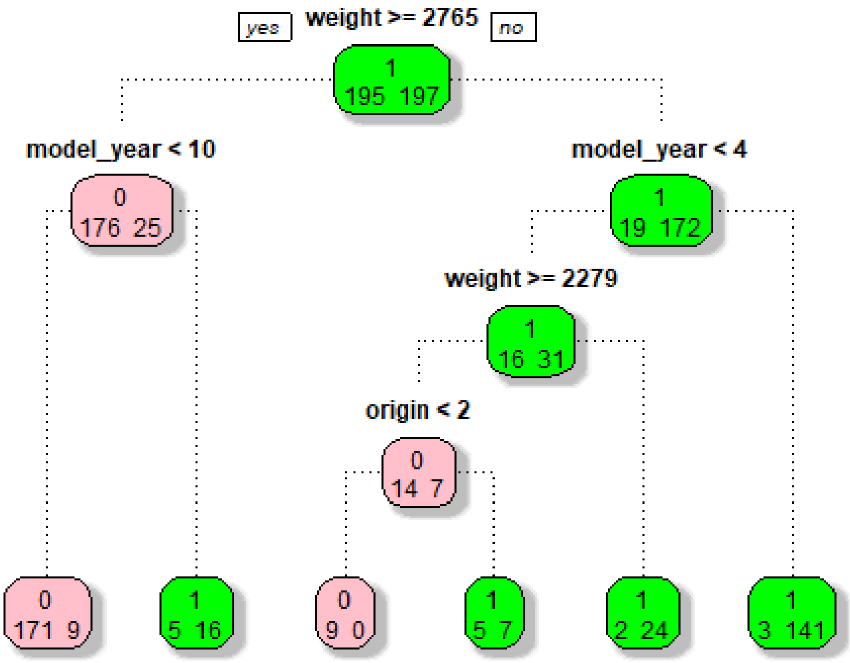
1. Clean dataset.
   - Model year has been changed to “0” for the year 1970, “1” for 1971, and so on.
   - Displacement was changed to the total volume of the engine(cylinder times original displacement numbers).
2. Create a logistic regression(Binomial Model) model to identify which of the characteristics that might be statistically significant.
3. Create a decision tree with the same characteristics as the logistic model.
4. Compare the two models.
5. Conclusion

```
##
## Call:
## glm(formula = mpg_good ~ horsepower + weight + model_year + origin,
##     family = "binomial", data = autodf)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.2841  -0.1762   0.0334   0.2872   4.9667
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) 10.5482227  1.6374222   6.442 1.18e-10 ***
## horsepower  -0.0289223  0.0138262  -2.092   0.0365 *
## weight      -0.0037458  0.0005972  -6.272 3.55e-10 ***
## model_year   0.3597694  0.0628446   5.725 1.04e-08 ***
## origin       0.4246857  0.2687721   1.580   0.1141
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 543.42  on 391  degrees of freedom
## Residual deviance: 185.38  on 387  degrees of freedom
## AIC: 195.38
##
## Number of Fisher Scoring iterations: 7
```

Logistical business insights

- First off, I started with all the different variables that the dataset provided: Number of cylinders, displacement(size of each cylinder), horsepower, weight, acceleration, model year.
  - I kept trimming the model until all the models are statistically significant (except origin).
- Based on the numbers above, we can conclude the following:
  - If the number of horsepowers increases by 1, the probability of the car being environmental friendly decreases by 4.4 percent.
  - If the weight of the car increases by one pound, the probability of the vehicle being environmental friendly decreases by 0.43 percent.
  - The newer the car, the more environmental friendly. Increasing the model year by 1 will increase the probability of the car being environmental friendly by 53.6 percent.
  - It also looks like the European and Asian cars tends to be more environmental friendly than the American cars.

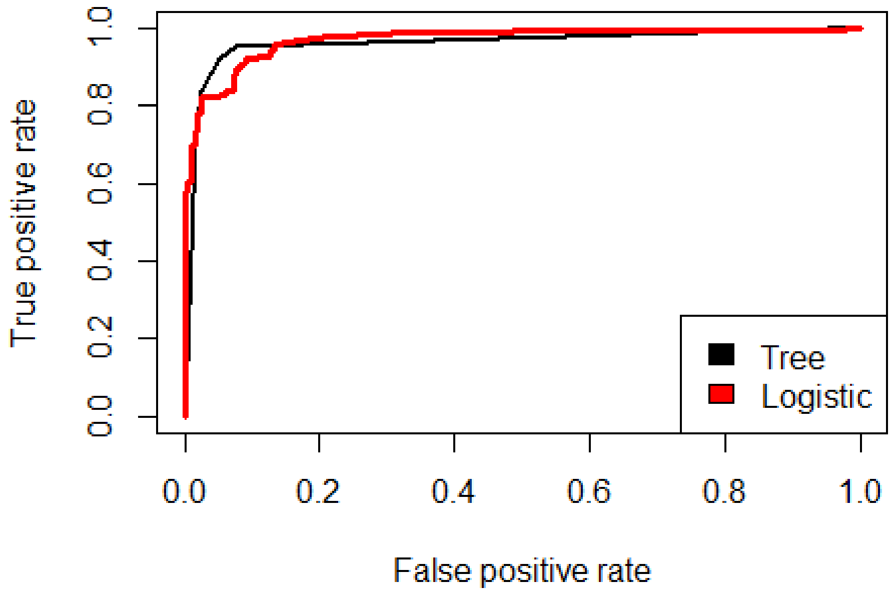
Now to the decision tree



Decision tree insight

- As mentioned earlier, I am using the same characteristics from the logistical regression.
- Here you are provided some examples to make it easier to read the tree:
  - A car that weighs more or equal to 2,765 pounds which was created before 1980, tends to be less environmental friendly (172 bad vs. 8 good).
  - A car that weighs more or equal to 2,765 pounds which was created after 1980, tends to be more environmental friendly (5 bad vs. 16 good).
  - A car that weighs less than 2,765 pounds which was created after 1974, tends to be more environmental friendly (3 bad vs. 141 good).
  - A car which weights less than 2,279 pounds which was created before 1974, tends to be more environmental friendly (2 bad vs. 24 good).

Let’s compare the two models in regards to performance.



Performance comparison

- Red line represents the logistic regression. The black line represents the decision tree.
- Both the model seem to be very similar, and there is hard to tell if one of them is stronger than the other one unless we look at some specific values (for example illustrated in the top left corner where the black graph is above the red).

Conclusion

To conclude, you can see that the logistical regression model takes more numbers into account; horsepower, weight, model year, and origin, whereas the tree only looks into weight, model year (and one branch of origin). Both of the models provide us the understanding that weight and which year the car was manufactured has a high impact on the mileage per gallon. Year makes a lot of sense after doing some more research since in 1975 the Congress passed a new law to increase fuel efficiency. For further information on this, you can click the following link: Driving to 54.5 MPG