Kelvyn Bladen

Stats 5120 – Dr. Cutler

# Project Report

## Data

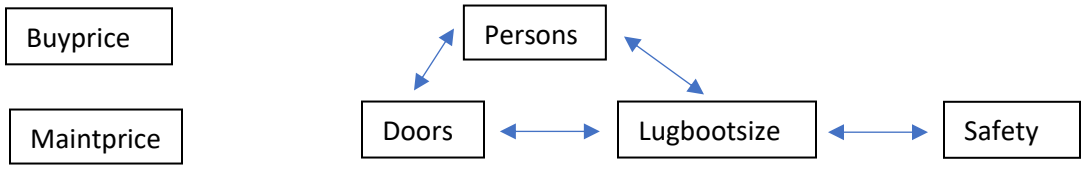
The data set that I choose for this project comes from the car evaluation sets in the UC Irvine Data Repository. The response variable is a rating for each car. These ratings are unacceptable, acceptable, good, and very good. From these ratings, I created numerical ratings of 1-4 for the response and built another response variable coded as 0 for unacceptable, and 1 for other. This data set has 6 predictor variables and 1728 observations. The predictor variables are: Buy Price, Maintenance Price, Doors, Persons, Safety, and Lug Boot Size. Each of these variables is categorical or discrete. Car and maintenance price have 4 categories: low, medium, high, and very high. Doors has discrete values of 2, 3, 4, and 5 or more, while the persons variable is coded similarly: 2, 4, and more. Lug boot size has 3 values: small, medium, and big. Safety also has 3 values: low, medium, and high. From this data, I built logistic regression models, as well as a few other classification models for understanding which factors are most influential in determining a car's rating.

## Logistic Regression

Using logistic regression, I created models for both of my response variables. I used a cumlogit link function to model the ratings, whereas for my binary response I used a standard logit link. For the first model, after trimming the insignificant interaction terms, I got a model with the following terms remaining:

| Type 3 Analysis of Effects | | | |
|---|---|---|---|
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **buyprice** | 3 | 2081.5961 | <.0001 |
| **maintprice** | 3 | 1932.6477 | <.0001 |
| **doors** | 3 | 0.0081 | 0.9998 |
| **persons** | 2 | 37.1085 | <.0001 |
| **doors*persons** | 6 | 491.6957 | <.0001 |
| **lugbootsize** | 2 | 0.0134 | 0.9933 |
| **doors*lugbootsize** | 6 | 785.8975 | <.0001 |
| **persons*lugbootsize** | 4 | 409.2505 | <.0001 |
| **safety** | 2 | 1735.4113 | <.0001 |
| **lugbootsize*safety** | 4 | 430.1375 | <.0001 |

This yields the following conditional independence plot:

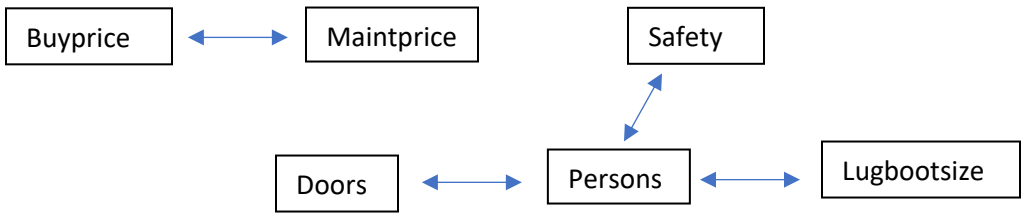



To evaluate the goodness of fit for this model, I used the genmod procedure in SAS with these same variables to find the scaled deviance of 427.5397. Comparing this to 5146 degrees of freedom, I'd say this model fits extremely well. In fact, the model seems to be suffering from some amount of overfitting.

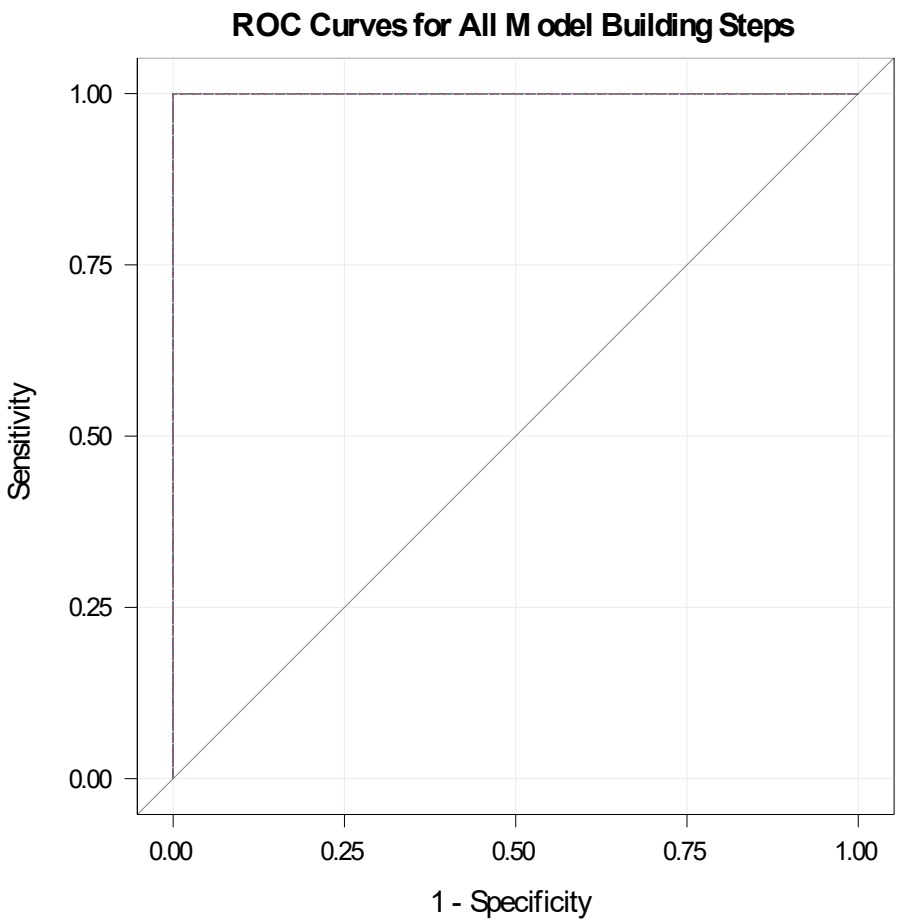
Looking at the results for the logistic regression for the binary response, I found a slightly different set of predictors:

| Type 3 Analysis of Effects | | | |
|---|---|---|---|
| **Effect** | **DF** | **Wald Chi-Square** | **Pr > ChiSq** |
| **buyprice** | 3 | 10096.7606 | <.0001 |
| **maintprice** | 3 | 9568.4172 | <.0001 |
| **buyprice*maintprice** | 9 | 7798.4788 | <.0001 |
| **doors** | 3 | 1921.4810 | <.0001 |
| **persons** | 2 | 3575.9810 | <.0001 |
| **doors*persons** | 6 | 5259.9127 | <.0001 |
| **lugbootsize** | 2 | 3880.0883 | <.0001 |
| **persons*lugbootsize** | 4 | 3782.8594 | <.0001 |
| **safety** | 2 | 7229.7930 | <.0001 |
| **persons*safety** | 4 | 5133.1759 | <.0001 |

The conditional independence for this set would therefore look as follows:




The ROC curve and c-table below tell the same story as the first model. For cutoff values between 0.1 and 0.9, the model is a perfect fit. Consequently, the model is extremely accurate, but I'm still seeing strong evidence of overfitting.



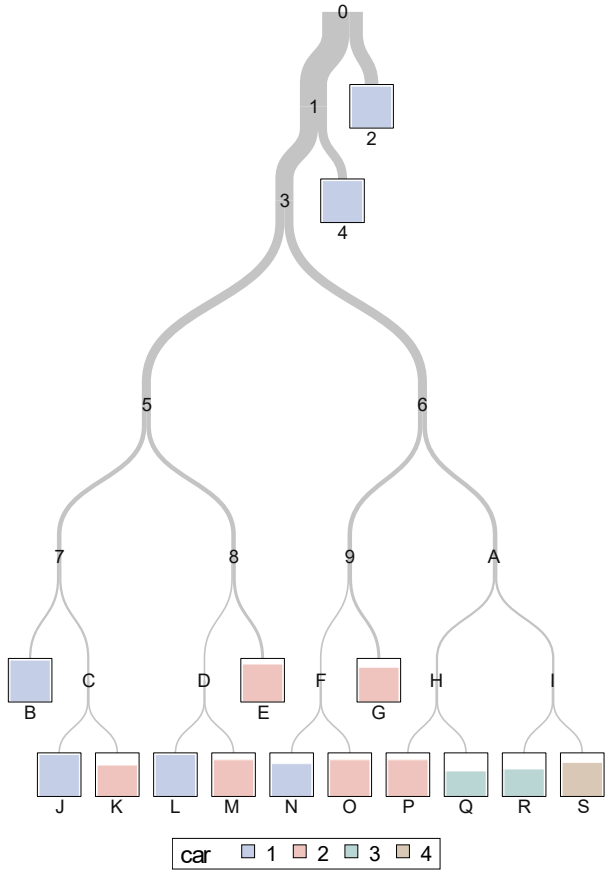

| Classification Table | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | Correct | | Incorrect | | Percentages | | | | |
| Prob Level | Event | Non-Event | Event | Non-Event | Correct | Sensi-tivity | Speci-ficity | Pos Pred | Neg Pred |
| **0.000** | 1210 | 0 | 518 | 0 | 70.0 | 100.0 | 0.0 | 70.0 | . |
| **0.100** | 1210 | 518 | 0 | 0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **0.500** | 1210 | 518 | 0 | 0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **0.900** | 1210 | 518 | 0 | 0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |
| **1.000** | 0 | 518 | 0 | 1210 | 30.0 | 0.0 | 100.0 | . | 30.0 |

**Classification Trees and Random Forest**

Due to the possible concern for overfitting, I used a few other types of analysis for understanding this dataset. From stat learning and data mining, I've built tree-based models and random forest algorithms for evaluating classification data such as this. My intent for creating these models is to understand which predictors are most important to a car's acceptability rating, rather than to generate greater predictive power.

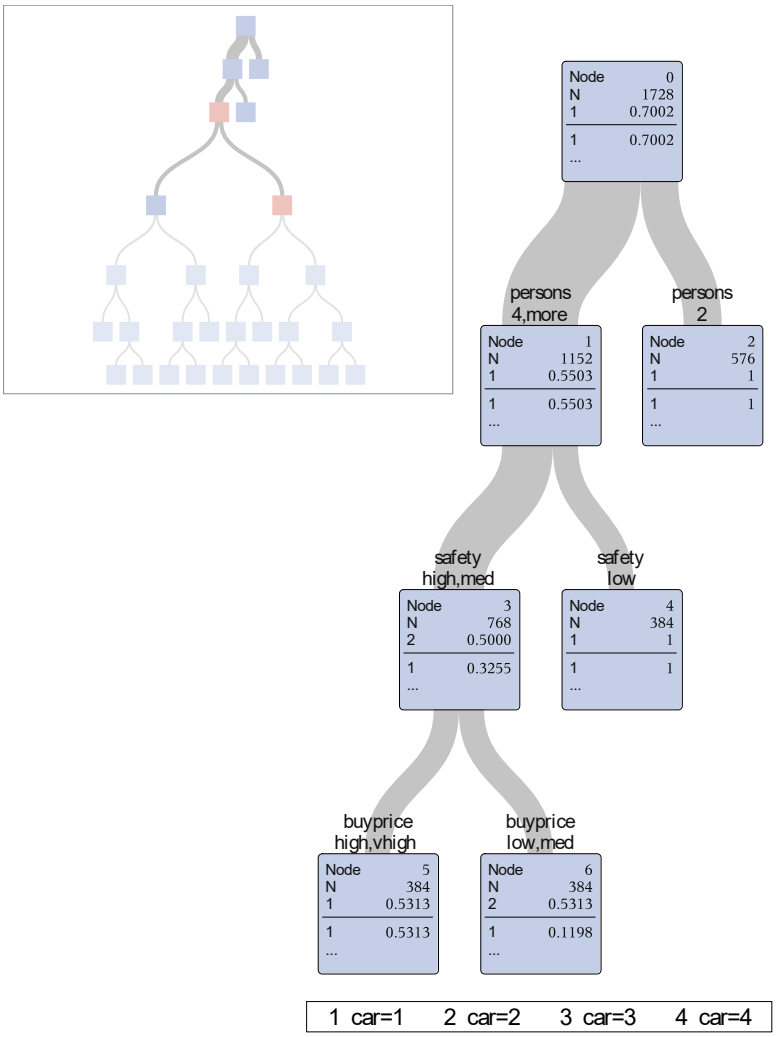
Starting with the classification tree, I ran a test to find how many terminal nodes I should use. I found that a 60-node tree was able to describe the data perfectly. However, this is extremely large and overly complicated. Consequently, after trying a couple of options, I decided that a 15-node tree provides a strong predictive model, without being overly complicated. The plot (below) indicates that the nodes appear to be quite pure, while the confusion matrix shows that the tree's cross-validated accuracy is about 93.11%. One benefit of a tree is that it helps to see where variables are splitting and which splits are most important. In this case, if persons = 2 or safety = "low", I would already be able to decide that the car should have a rating of unacceptable. I can also clearly see that only cars meeting the criterion for node A should have ratings of good or very good. Like a dichotomous key, I can quickly check different attributes and follow them down the tree to the corresponding node. If they lead to node M, I would confidently predict the car to be acceptable. If it matches node Q, I would expect the car to have a good rating.  After building this tree, SAS provides an idea of which variables are most important with a small summary report (below). From this, we see that safety, then persons, are most important in determining a car's rating. The variable, doors, was never used for splitting, so I'd expect that it is the least relevant to a car's rating.

**Classification Tree for car**




| Confusion Matrices | | | | | | |
|---|---|---|---|---|---|---|
| | | Predicted | | | | Error |
| | Actual | 1 | 2 | 3 | 4 | Rate |
| **Model Based** | **1** | 1161 | 45 | 4 | 0 | 0.0405 |
| | **2** | 7 | 342 | 32 | 3 | 0.1094 |
| | **3** | 0 | 0 | 60 | 9 | 0.1304 |
| | **4** | 0 | 13 | 0 | 52 | 0.2000 |
| **Cross Validation** | **1** | 1161 | 45 | 4 | 0 | 0.0405 |
| | **2** | 7 | 343 | 31 | 3 | 0.1068 |
| | **3** | 0 | 7 | 53 | 9 | 0.2319 |
| | **4** | 0 | 13 | 0 | 52 | 0.2000 |

## Subtree Starting at Node=0




| Variable Importance | | | |
|---|---|---|---|
| | Training | | |
| Variable | Relative | Importance | Count |
| safety | 1.0000 | 15.7091 | 4 |
| persons | 0.7057 | 11.0863 | 1 |
| maintprice | 0.7040 | 11.0588 | 3 |

| Variable Importance | | | |
|---|---|---|---|
| | Training | | |
| Variable | Relative | Importance | Count |
| **lugbootsize** | 0.5044 | 7.9244 | 4 |
| **buyprice** | 0.4867 | 7.6456 | 2 |

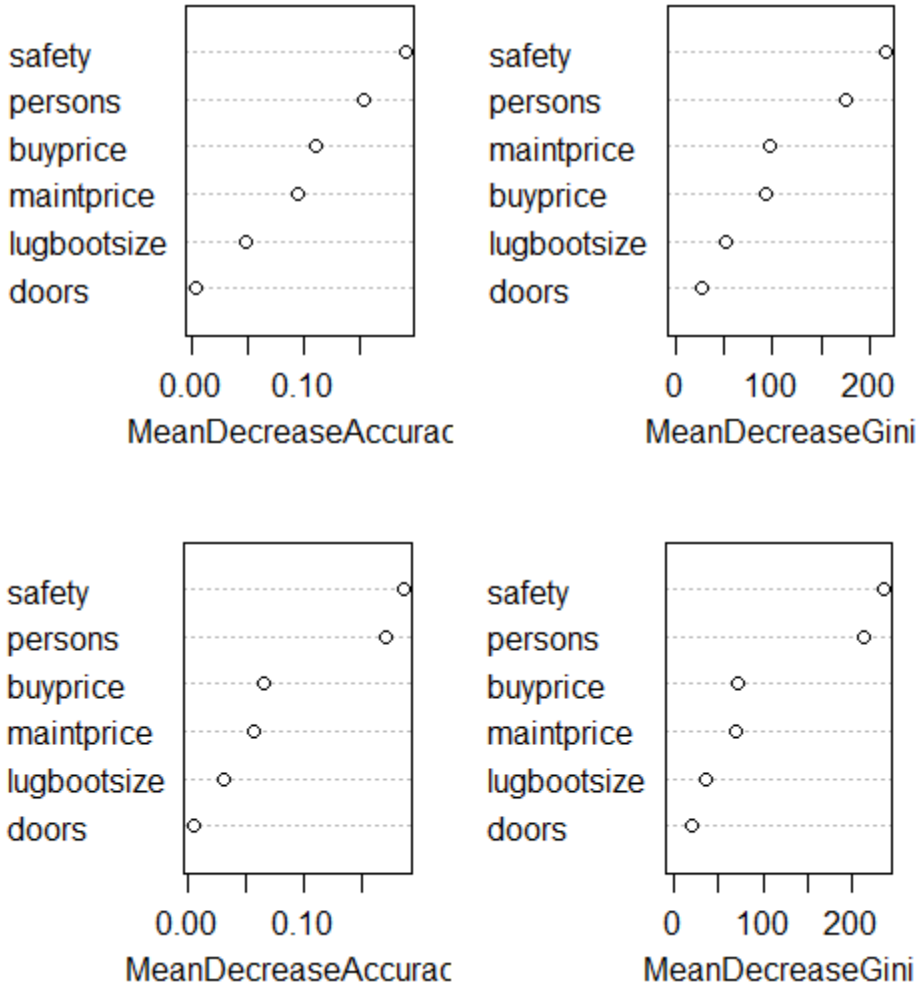
In R, I then ran my data through the random forest algorithm for both of my response variables to continue this analysis. From the confusion matrix and percent correctly classified, I can see this method has an error of 2.154% for my first response.

```
     1   2  3  4 class.error
1 1195  14  1  0  0.01239669
2    3 375  3  3  0.02343750
3    0   3 60  6  0.13043478
4    0   4  0 61  0.06153846

PCC: 97.8459
```

The accuracy for the model of the binary response is comparable (see below). Naturally, the error is a bit smaller at 1.16%.

```
     0   1 class.error
0 1193  17 0.014049587
1    3 515 0.005791506


PCC: 98.84146
```

However, the main reason for doing this was to use random forest to evaluate variable importance, rather than build strong confusion matrices. The charts below show the same order of variable importance for both response variables.




**Conclusion**

From the random forest analysis, we can see that across the board, safety has the highest relevance to a car's acceptability rating. Next is persons, followed by buyprice and maintprice, which are quite comparable in their influence. Finally, lugbootsize and doors have the least bearing on rating. Comparing these two plots to the one from the classification tree, we see some close similarities. However, the random forest plots give higher emphasis to buyprice and lower emphasis to lugbootsize. Since random forest is a boot-strapping technique involving hundreds of classification trees, I tend to prefer its variable importance results over those generated by a single tree.

Returning to my logistic regression results, the perfect fit yields some exotic odds ratios. Just as an example, 2 people and low safety is about 4 million times more likely to be unacceptable than low safety and 5 or more people. Also, 4 people and high safety is 1678 times more likely to be acceptable than medium safety and 4 people. Similar results of this magnitude can be found by exponentiating the estimated coefficients from the logistic regression. Consequently, the provided odds ratios don't tell us much except direction. A better way to understand this data is by looking at the classification tree for general trends rather than specific odds. While it's difficult to quantify all these differences, the graphical evidence is crystal clear. When considering buyprice, the tree shows that lower prices are much more likely to be acceptable than higher ones. The same principle is true for maintprice. When comparing safety, low safety is least acceptable, while high safety is more likely to be rated very good. For lugbootsize, small sizes aren't as good for car ratings as medium or large sizes. Finally, two person cars are much worse for ratings than cars that fit 4 or more.