Car Evaluation

Sravani Chatrathi

1. Introduction

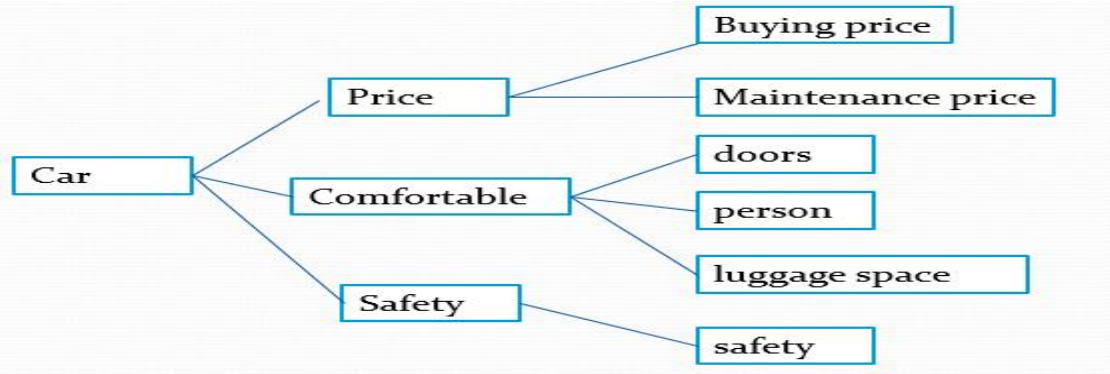
Cars are the most common means of transport in the modern-day world. Almost every person aged between 18-65 years has a car or will be aspiring to buy a car. Car purchasing involves a lot of factors. This dataset contains all those factors which helps in deciding which factors play a key role in choosing a car.

This evaluation gives the Car Manufacturing companies an idea on what the users are looking for. This also helps them to design cars emphasizing on features that lead to car acceptability.

In this report, I have performed Supervised classification on the Car evaluation dataset based on various predictors like buying price, maintenance price, estimated safety, number of persons, number of doors, luggage space. Best subset selection was employed for feature selection. Supervised Classification was performed on the dataset using the Logistic Regression. A validation set approach method is used on test dataset and the misclassification rate is computed. When the predictors persons, buying, safety, maint, lug_boot, doors were considered, an accuracy of 0.948 was achieved.

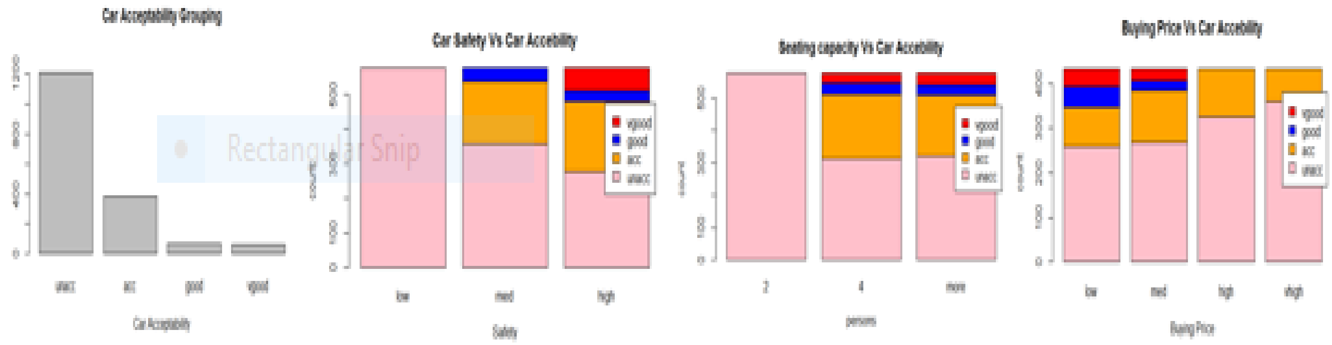
2. Dataset

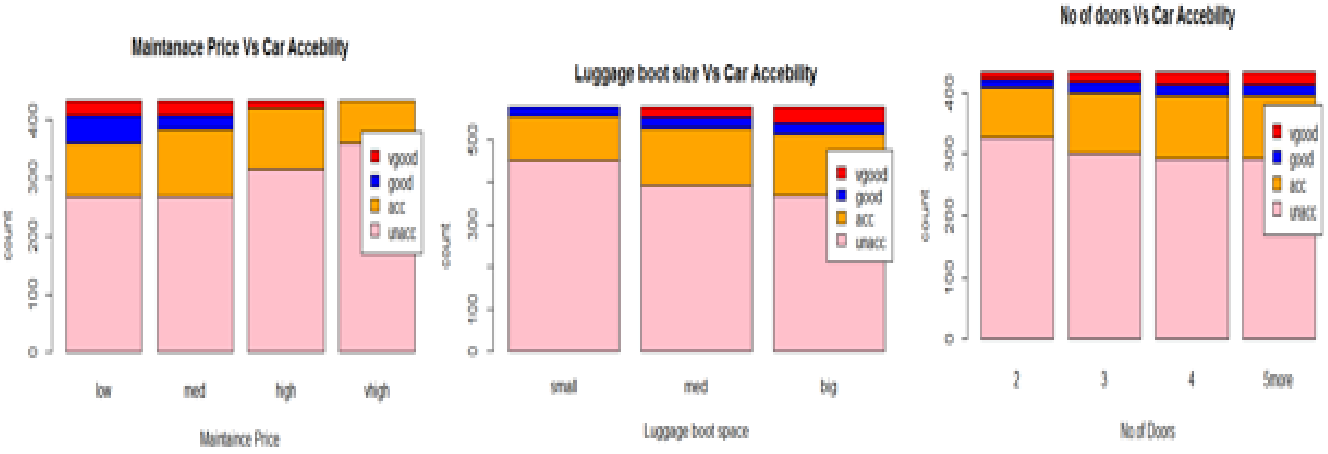
This dataset that was derived from a simple hierarchical decision model originally for developed model originally developed for the demonstration of DEX (M. Bohanec, V. Rajkovic: Expert system for decision making. Sistemica 1(1), pp. 145-157, 1990.). There are 1728 observations and 7 variables. It contains 6 Predictors and 1 response variable. The variable car acceptability is the outcome and it has four classes *vgood, good, acc, unacc.* The 6 predictors are namely buying price *(vhigh, high, med, low)*, price of maintenance *(vhigh, high ,med, low)*, number of doors(2,3,4,5more), Persons capacity in terms of persons to carry *(2, 4, more)*,Size of luggage boot *(small, med, big)*,Estimated safety of the car *(low, med, high)*



3. Exploratory Data Analysis

To understand the predictors, bar plots are plotted to examine the importance of each predictor in accepting or rejecting a Car. We can observe that safety and Seating Capacity are important factors in accepting or rejecting a car. Low safety is not accepted by customers.

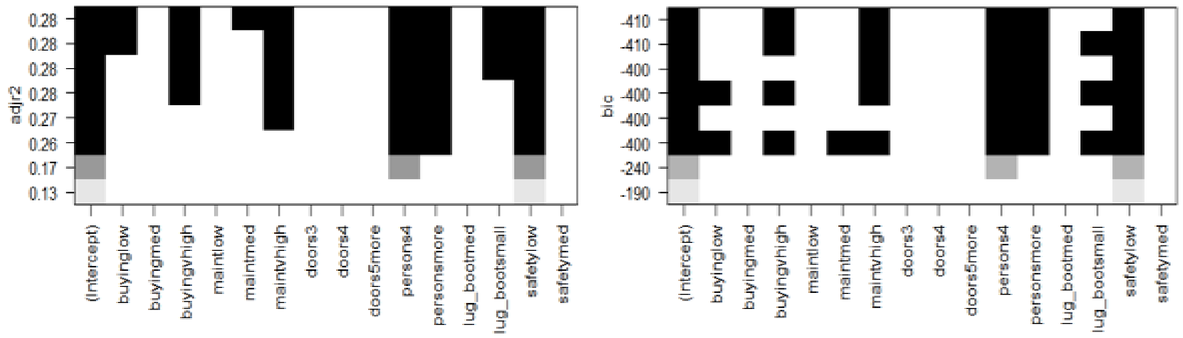




4. Model Selection and Validation

This dataset is split into 80% for training and 20% for testing purposes. Best subset selection was used for feature selection. Six predictors were considered for the analysis of the Supervised classification. The response variable Car acceptability contains 4 classes, so we need to use a Multinomial Classification Model. So, I have selected Multinomial Logistic Regression. R has a function multinom() which is used for classifying multiple output classes. The model is then tested on "test dataset" and a confusion matrix was constructed to assess the accuracy and also misclassification rate was computed.

Best Subset : The Best subset feature selection was used to find the most important predictors in the The Bestsubset regression selects the smallest subset that fulfills criteria. The output after performing the Bestsubset Regression is shown below. From the plot it is clear that doors are of no importance in evaluation of Car. It is also observed that safety,persons are most important features in the evaluation.



Model Prediction:

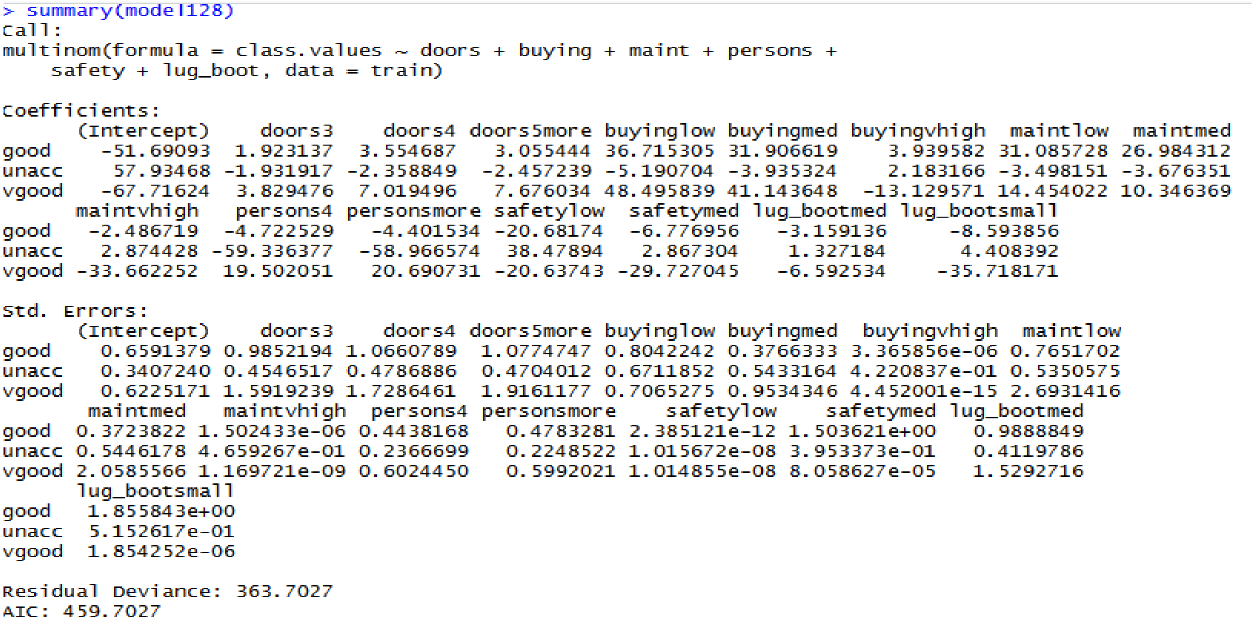
The Logistic regression was performed for each feature separately to understand its individual importance in the prediction. Then the predictors are added and their corresponding Accuracies and Misclassification Rate are tabulated below.

Predictors	**Accuracy**	**Misclassification Rate**
Safety	0. 7283	0. 2716
Persons	0. 7283	0. 2716
Doors	0. 7283	0. 2716
Buying	0. 7283	0. 2716
Maintenance	0. 7283	0. 2716
Luggage space	0. 7283	0.2716
Safety, Persons	**0.8064**	**0.1936**
Luggage space, Safety, Persons	0.841	0.1589
Doors, Safety, Persons	0.7977	0.2023

Maintenance, Safety, Persons	0.8179	0.1820
Buying, Safety, Persons	0.8237	0.1763
Buying, Safety, Persons, Maintenance	**0.8844**	**0.1156**
Buying, Safety, Persons, Luggage space	0.8613	0.187
Buying, Safety, Persons, Maintenance, Luggage Space	**0.9393**	**0. 0606**
Buying, Safety, Persons, Maintenance, Luggage space, Doors	**0.9451**	**0.0549**

From the above information, It is clear that almost all predictors are important. We also can observe that when Buying, Safety, Persons, Maintenance, Luggage space, doors features accuracy is the highest i.e 0.9480. The misclassification rate which was computed through the validation set approach is 0.0520 which is also the least when compared to others.

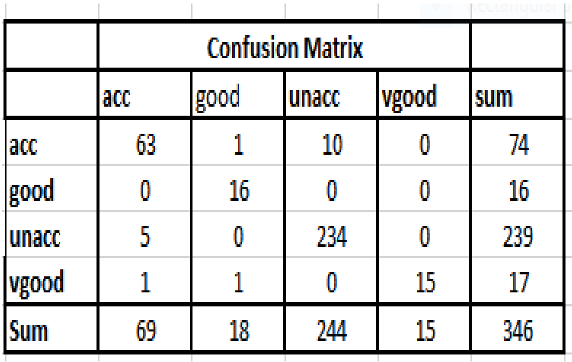
The below snippet gives the Summary of our Predicted model. It is also observed that the AIC and Residual Deviance are least when all the predictors are considered.



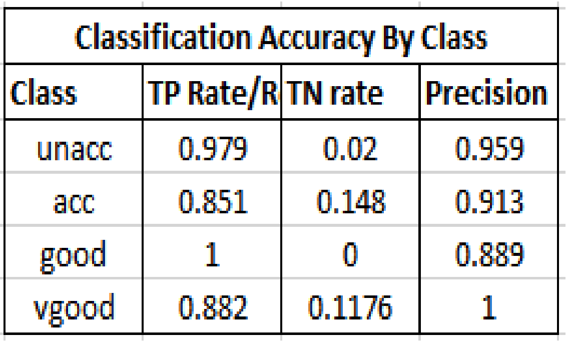
```
> summary(model128)
Call:
multinom(formula = class.values ~ doors + buying + maint + persons + 
    safety + lug_boot, data = train)

Coefficients:
      (Intercept)    doors3    doors4 doors5more buyinglow buyingmed buyingvhigh  maintlow  maintmed
good    -51.69093  1.923137  3.554687   3.055444 36.715305 31.906619    3.939582 31.085728 26.984312
unacc    57.93468 -1.931917 -2.358849  -2.457239 -5.190704 -3.935324    2.183166 -3.498151 -3.676351
vgood   -67.71624  3.829476  7.019496   7.676034 48.495839 41.143648  -13.129571 14.454022 10.346369
      maintvhigh   persons4 personsmore safetylow  safetymed lug_bootmed lug_bootsmall
good   -2.486719  -4.722529   -4.401534 -20.68174  -6.776956   -3.159136     -8.593856
unacc   2.874428 -59.336377  -58.966574  38.47894   2.867304    1.327184      4.408392
vgood -33.662252  19.502051   20.690731 -20.63743 -29.727045   -6.592534    -35.718171

Std. Errors:
      (Intercept)    doors3    doors4 doors5more buyinglow buyingmed  buyingvhigh  maintlow
good    0.6591379 0.9852194 1.0660789  1.0774747 0.8042242 0.3766333 3.365856e-06 0.7651702
unacc   0.3407240 0.4546517 0.4786886  0.4704012 0.6711852 0.5433164 4.220837e-01 0.5350575
vgood   0.6225171 1.5919239 1.7286461  1.9161177 0.7065275 0.9534346 4.452001e-15 2.6931416
       maintmed   maintvhigh  persons4 personsmore    safetylow    safetymed lug_bootmed
good  0.3723822 1.502433e-06 0.4438168   0.4783281 2.385121e-12 1.503621e+00   0.9888849
unacc 0.5446178 4.659267e-01 0.2366699   0.2248522 1.015672e-08 3.953373e-01   0.4119786
vgood 2.0585566 1.169721e-09 0.6024450   0.5992021 1.014855e-08 8.058627e-05   1.5292716
      lug_bootsmall
good   1.855843e+00
unacc  5.152617e-01
vgood  1.854252e-06

Residual Deviance: 363.7027 
AIC: 459.7027 
```

The residual deviance here is the error left in the model. It is similar to the sum of squares in Linear regression model.

In the above model acc is considered as a reference since the class.values(Car acceptability) is a categorical variable. The below tables provides details about the performance of our model :



	Confusion Matrix				
	acc	good	unacc	vgood	sum
acc	63	1	10	0	74
good	0	16	0	0	16
unacc	5	0	234	0	239
vgood	1	1	0	15	17
Sum	69	18	244	15	346



Classification Accuracy By Class			
Class	TP Rate/R	TN rate	Precision
unacc	0.979	0.02	0.959
acc	0.851	0.148	0.913
good	1	0	0.889
vgood	0.882	0.1176	1

5. Conclusion

Supervised classification was performed on several predictors to assess the accuracy and the key predictors which contribute to the Car Evaluation. It is inferenced that:

- All the variables are important for customers in assessing whether the car is in acceptable or unacceptable range
- Safety and Seating capacity are two main factors in rejecting the cars as unacceptable
- If the value of "safety" is low, the result will directly fall into unacceptable (unacc) and whatever the value of safety is, if Seating value is 2, the entry will also fall directly into unacceptable.
- Number of doors are the least important variable in deciding the class value of the car.