Car Evaluation Dataset

MANJANG Amadou, OSONDU Ikechi Chimereze, Roosvel Tatsinkou Tenekeu

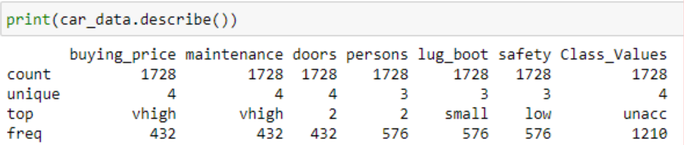
Data Analytics and Data Driven Decision
DISIM
University of L'Aquila, Italy

July 2023

Contents

1. Dataset
2. Description
3. data cleaning (if needed)
4. exploratory analysis
5. main analysis: objective and methods adopted
6. Decision Tree Classification
7. SUPPORT VECTOR CLASSIFICATION
8. LOGISTIC REGRESSION
9. RANDOM FOREST CLASSIFICATION
10. preview/summary of the results
11. Conclusion
12. Reference

Dataset I



```
print(car_data.describe())
```

```
       buying_price maintenance doors persons lug_boot safety Class_Values
count          1728        1728  1728    1728     1728   1728         1728
unique            4           4     4       3        3      3            4
top           vhigh       vhigh     2       2    small    low        unacc
freq            432         432   432     576      576    576         1210
```

- Dimensions : 1728 Rows, 7 Columns
- car evaluation data for customers to make decisions on which car to buy
- Target feature: Class values
- all cars are classified as either unacc, acc, good or very good

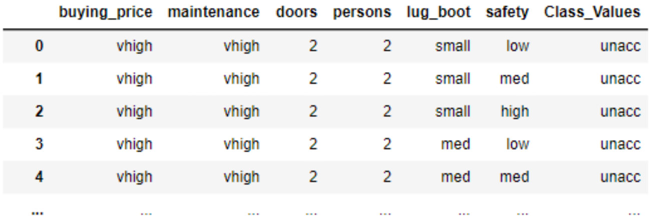
Description I

The car Evaluation dataset is mostly use for classification. The model evaluates cars according to the following concept structure:

- overall price
- buying price
- price of the maintenance
- number of doors
- capacity in terms of persons to carry
- the size of luggage boot
- estimated safety of the car

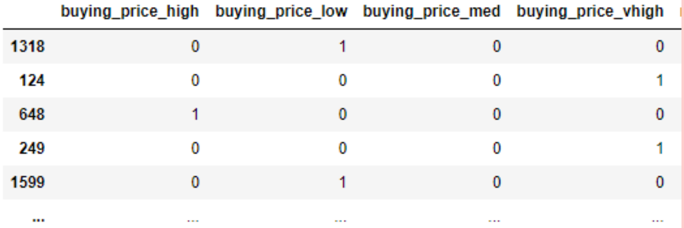
These features can be used by customers to make informed decisions on which cars to buy.

Data Preprocessing I



	buying_price	maintenance	doors	persons	lug_boot	safety	Class_Values
0	vhigh	vhigh	2	2	small	low	unacc
1	vhigh	vhigh	2	2	small	med	unacc
2	vhigh	vhigh	2	2	small	high	unacc
3	vhigh	vhigh	2	2	med	low	unacc
4	vhigh	vhigh	2	2	med	med	unacc
...	...	...	...	...	...	...	...

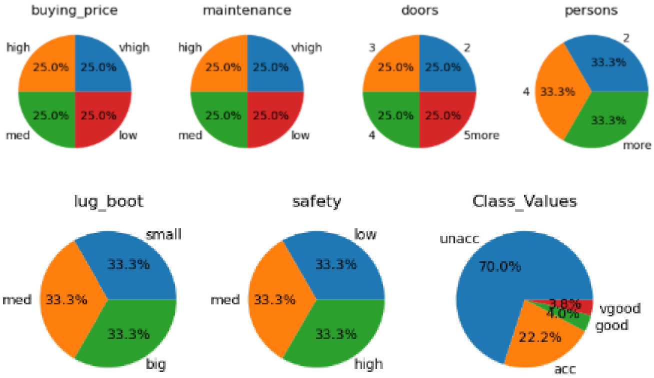
- No missing values
- We rename our columns to fit our needs



	buying_price_high	buying_price_low	buying_price_med	buying_price_vhigh
1318	0	1	0	0
124	0	0	0	1
648	1	0	0	0
249	0	0	0	1
1599	0	1	0	0
...	...	...	...	...

- Our data is all categorical
- we use One-hot Encoding to turn our data in to numerical data
- Split in to Training and Test data at 80 to 20 percent ratio
- Set out Target variable to Class values

Pie Chart I



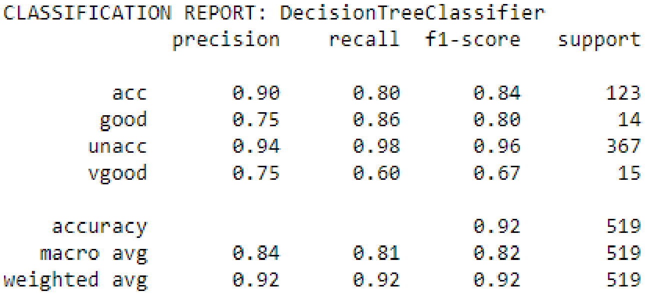
- it could be seen from the charts that our categories for each feature have equal number of instances
- therefore our data is a balanced dataset.
- it is sufficient to use accuracy to compare the performance of our models

Supervised Learning

Having explored, processed and analysed the dataset, we are now ready to use the data for training and testing a model that can be used by customers and sails person to make decisions on the purchase of a car given some specific features/characteristics. In this light we used four Supervised learning classification algorithms:

- Decision tree classification,
- Support vector classification,
- Logistic regression and
- Random forest classification classification.

Decision Tree Classification



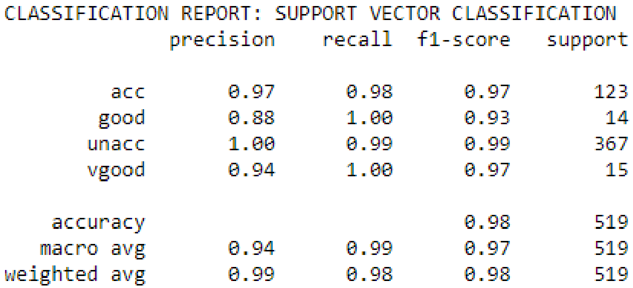
```
CLASSIFICATION REPORT: DecisionTreeClassifier
              precision    recall  f1-score   support

         acc       0.90      0.80      0.84       123
        good       0.75      0.86      0.80        14
       unacc       0.94      0.98      0.96       367
       vgood       0.75      0.60      0.67        15

    accuracy                           0.92       519
   macro avg       0.84      0.81      0.82       519
weighted avg       0.92      0.92      0.92       519
```

Considering that the model has a high precision, recall and f1-score for all the classes. we can conclude that the model did a very good job at predicting the data.

SUPPORT VECTOR CLASSIFICATION



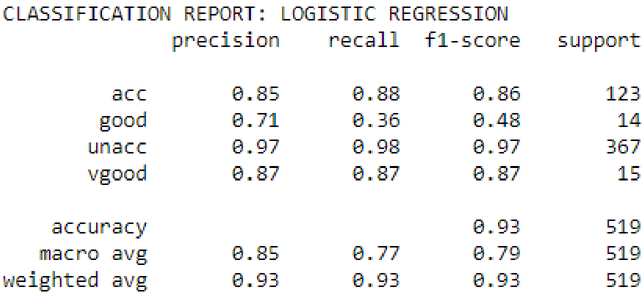
```
CLASSIFICATION REPORT: SUPPORT VECTOR CLASSIFICATION
              precision    recall  f1-score   support

         acc       0.97      0.98      0.97       123
        good       0.88      1.00      0.93        14
       unacc       1.00      0.99      0.99       367
       vgood       0.94      1.00      0.97        15

    accuracy                           0.98       519
   macro avg       0.94      0.99      0.97       519
weighted avg       0.99      0.98      0.98       519
```

Considering that the model has a high precision, recall and f1-score for all the classes. we can conclude that the model did an excellent job at predicting the data.

LOGISTIC REGRESSION



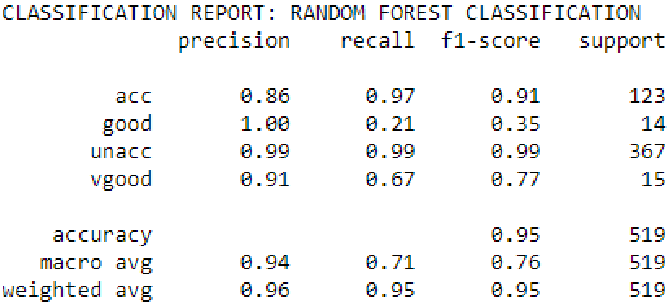
```
CLASSIFICATION REPORT: LOGISTIC REGRESSION
              precision    recall  f1-score   support

         acc       0.85      0.88      0.86       123
        good       0.71      0.36      0.48        14
       unacc       0.97      0.98      0.97       367
       vgood       0.87      0.87      0.87        15

    accuracy                           0.93       519
   macro avg       0.85      0.77      0.79       519
weighted avg       0.93      0.93      0.93       519
```

The model predicts the good class with moderate precision but only capture 36 percent of the good class in the data. An f1 score of 0.48 suggests moderate balance between the precision and the recall.

RANDOM FOREST CLASSIFICATION



```
CLASSIFICATION REPORT: RANDOM FOREST CLASSIFICATION
              precision    recall  f1-score   support

         acc       0.86      0.97      0.91       123
        good       1.00      0.21      0.35        14
       unacc       0.99      0.99      0.99       367
       vgood       0.91      0.67      0.77        15

    accuracy                           0.95       519
   macro avg       0.94      0.71      0.76       519
weighted avg       0.96      0.95      0.95       519
```

The model predicts the instances accurately. But the model was only able to capture 21 percent of the 'good' instances in the dataset. An f1 score 0.35 indicates that this model is not optimal in balancing precision and recall.

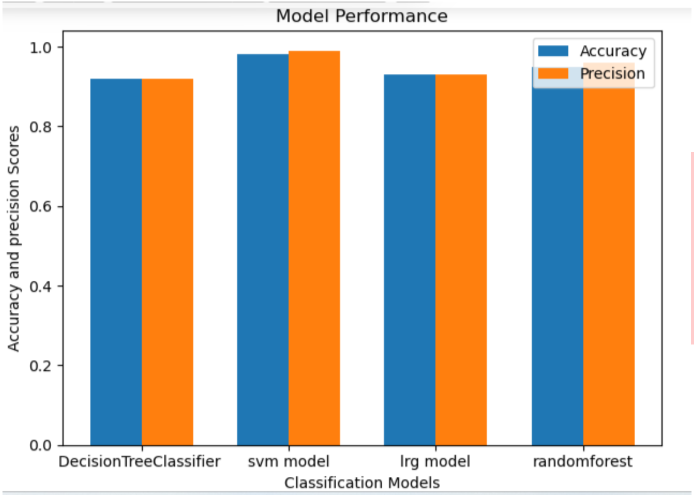
Results I

Table 1: Model Evaluation results

Model/performance	DecisionTreeClassifier	svm$_{m}odel$	lrg$_{m}odel$	randomforest
Accuracy	0.92	0.98	0.93	0.95
Precision	0.92	0.99	0.93	0.96
Recall	0.92	0.98	0.93	0.95
F1-score	0.92	0.98	0.93	0.95

Overall we can see that the support vector machine is the best model among the rest with an Accuracy of 0.98, precision 0.99, recall 0.98 and f1-score of 0.98. the second best model is the random forest, followed by logistic regression and the decision trees.

Model performance Bar Chart



The bar chart also confirms our claims from the table that the support vector classification out performs all the models. followed by random forest classification

Result

- Cars with low safety are considered unacceptable
- Cars that can take only 2 people are also unacceptable
- Cars with less than very high maintenance, with high safety, a boot that is not small and low buying price are very good cars
- Cars with high safety rating generally recieve above unacceptable rating
- No good or very good cars with very high maintenance cost
- Cheap cars generally are at least acceptable
- The number of doors have the least effect on our classification

Conclusion I

- Implemented decision tree, support vector classification, logistic regression and random forest models
- Evaluated evaluation metrics(Precision, Recall, Accuracy and F1 score).
- Support vector classification performed slightly better out of the four models.

References

- Bohanec,Marko. (1997). Car Evaluation. UCI Machine Learning Repository. https://doi.org/10.24432/C5JP48.