Data Mining & Processing

by Yusif Imamverdiyev
Rana Najafzade
Natavan Mammadova

MSc DSAI & MSc PCCE 2021

Agenda

- Problem Description
- Apriori SPMF
- Random Forest Algorithm
- Data Visualization and Processing for Modeling

Problem Description

- Anticipating the target values (column Class).
- Forecasting whether or not a car would be accepted.

Algorithms

Apriori SPMF

The process of locating frequent

itemset is depicted in as follows:

TID	items
T1	I1, I2 , I5
T2	I2,I4
T3	I2,I3
T4	I1,I2,I4
T5	I1,I3
T6	I2,I3
T7	I1,I3
T8	I1,I2,I3,I5
T9	I1,I2,I3

minimum support count is 2
minimum confidence is 60%

Itemset	sup_count
I1,I2	4
I1,I3	4
I1,I5	2
I2,I3	4
I2,I4	2
I2,I5	2
I2,I5	2

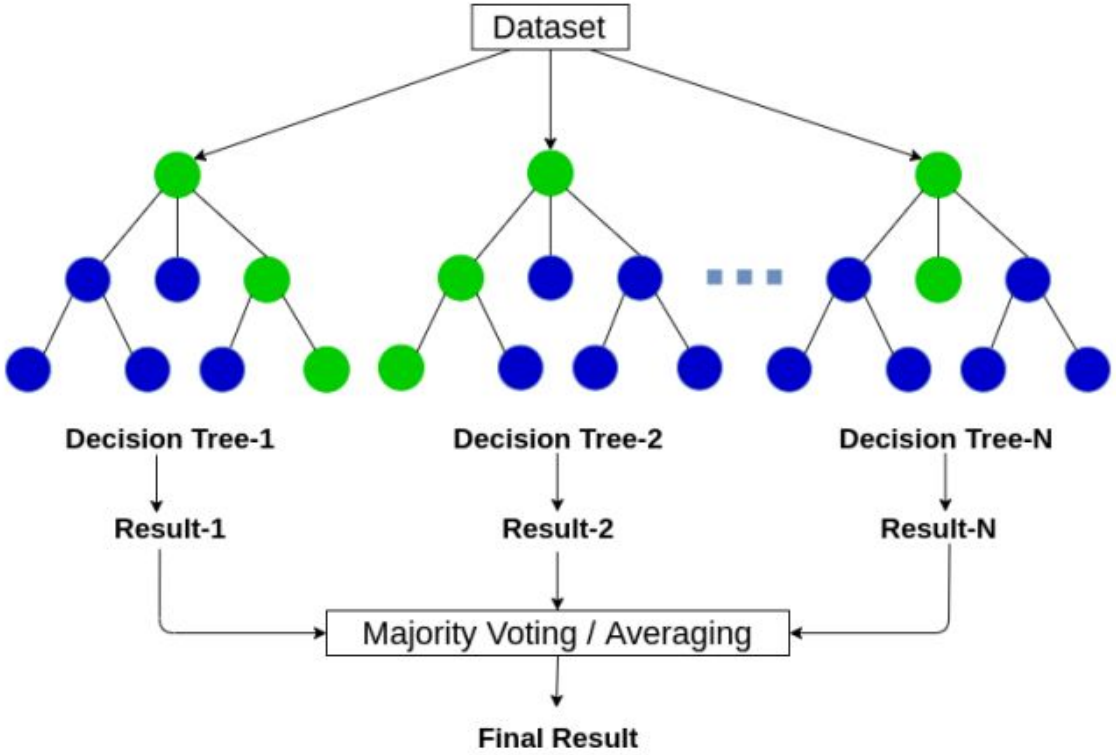
k = 2

Itemset	sup_count
I1,I2,I3	2
I1,I2,I5	2

k = 3

Algorithms

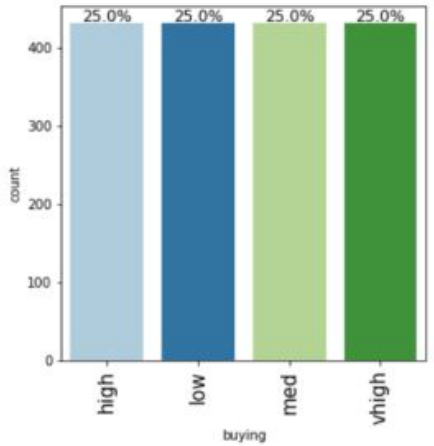
Random Forest



Data Visualization and Processing for Modeling

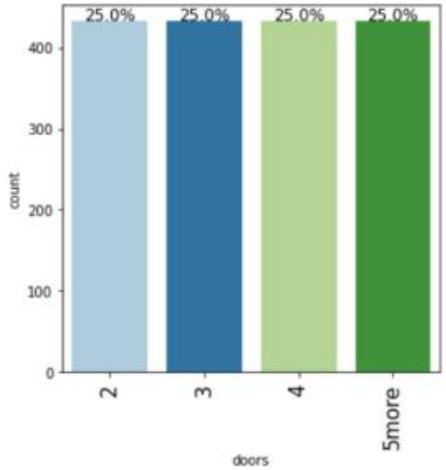
Observation on buying

```
In [51]: labeled_barplot(df, "buying", perc=True)
```



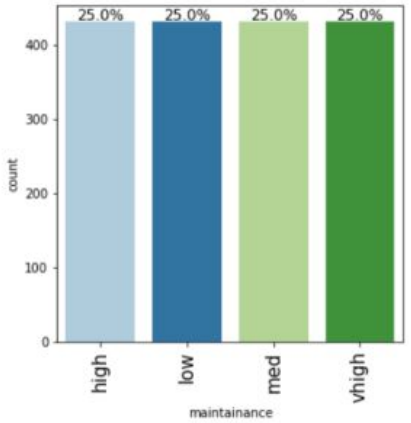
Observation on doors

```
In [53]: labeled_barplot(df, "doors", perc=True)
```



Observation on maintainance

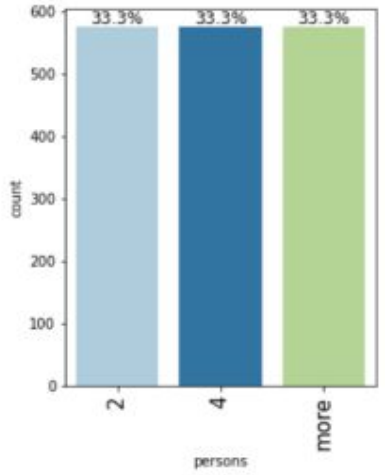
```
2]: labeled_barplot(df, "maintainance", perc=True)
```



Data Visualization and Processing for Modeling

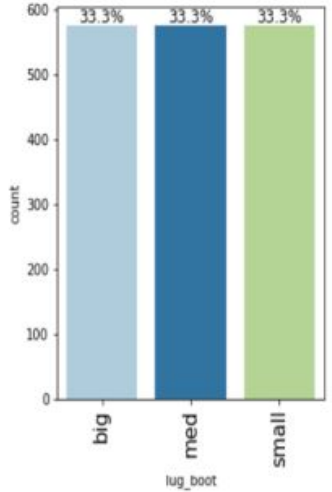
Observation on persons

```
In [54]: labeled_barplot(df, "persons", perc=True)
```



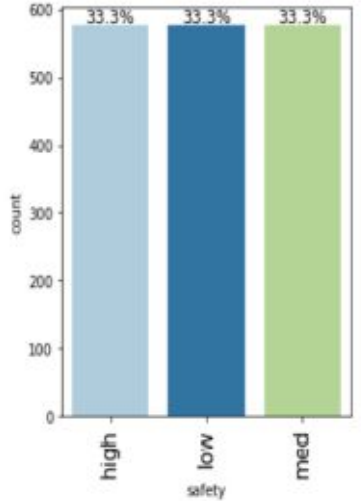
Observation on lug_boot

```
In [55]: labeled_barplot(df, "lug_boot", perc=True)
```



Observation on safety

```
In [56]: labeled_barplot(df, "safety", perc=True)
```



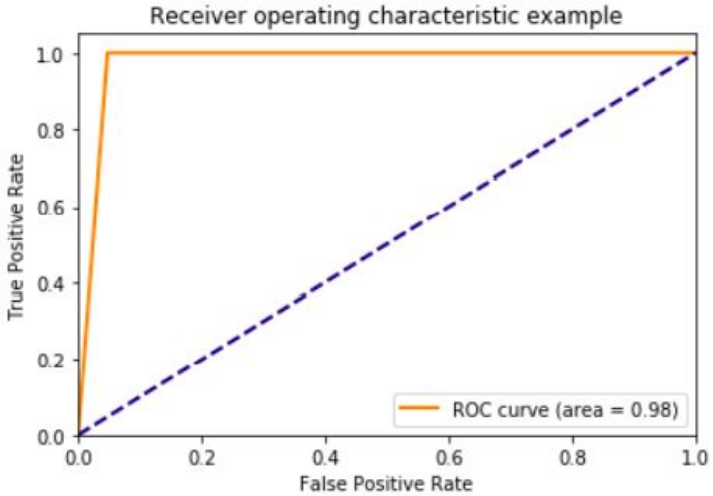
After bi variate analysis, we can see that if the safety is low and number of persons are 2 there is more chance for the car to be unacceptable.

```
Before UnderSampling, counts of label '1': 375
Before UnderSampling, counts of label '0': 834

After UnderSampling, counts of label '1': 375
After UnderSampling, counts of label '0': 375

After UnderSampling, the shape of train_X: (750, 21)
After UnderSampling, the shape of train_y: (750,)
```

undersample the data to balance the classes

Our conf matrix

```
Sensivity = 0.952247191011236
Specificity = 1.0
```



	precision	recall	f1-score	support
0	1.00	0.95	0.98	356
1	0.91	1.00	0.95	163
accuracy			0.97	519
macro avg	0.95	0.98	0.96	519
weighted avg	0.97	0.97	0.97	519

And other metrics