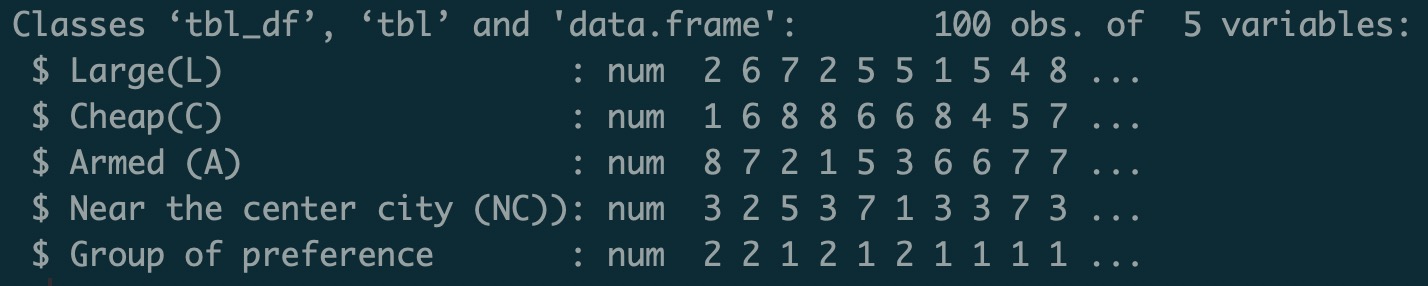
Oleksandr Romanchenko

83459

Decision tree

First, loading the data and checking its structure: whether all columns are numeric



Renaming columns and checking how many observations were classified to each group

A picture containing device, meter

Description automatically generated

Dividing data on training and validation datasets

A close up of a sign

Description automatically generated

Building first tree model

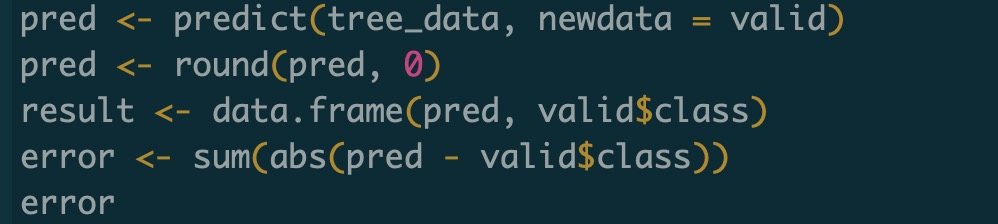


Getting the plot with tree

A close up of a logo

Description automatically generated

Now making prediction for validation dataset



A close up of a logo

Description automatically generated

As we can see – model predicts 6 cases incorrectly.

Now let’s try **rpart** model with the same data:



Setting up complexity parameter to a very small number I can increase model complexity.

Let’s have a look at the plot which shows how the error decreases when **cp** increases

A close up of a map

Description automatically generated

Draw a plot with tree

A screenshot of a cell phone

Description automatically generated

Plotting fancy version of tree diagram using **rattle** package

A close up of a map

Description automatically generated

Plot shows that:

* at the 1st split condition (cost>=3.5) was fulfilled for 47 cases and wasn’t for 23. These 23 were all classified to class 2 (because average number is 2 – meaning that all numbers there = 2)
* at the 2nd split condition (nearcity >=2.5) was fulfilled for 37 cases and wasn’t for 10. These 10 were all classified to class 2
* at the 3rd split condition (fited >= 3.5) was fulfilled for 29 cases and wasn’t for 8.

These 29 were all classified to class 1. For 8 cases leaf shows average 1.4 which tells that some of them were classified to class 1 and some to class 2

Now, let’s check prediction with this model and compute the error

A close up of a logo

Description automatically generated

A close up of a logo

Description automatically generated

The model predicts all cases in validation dataset correctly

**Conclusion**

During the experiment I’ve tried 2 models: **tree** and **rpart**. Tree model did 6 mistakes on validation data, rpart – 0 mistakes. In previous homework I’ve used neural network which did 1 mistake so I can conclude that rpart model performed even a bit better than network.