

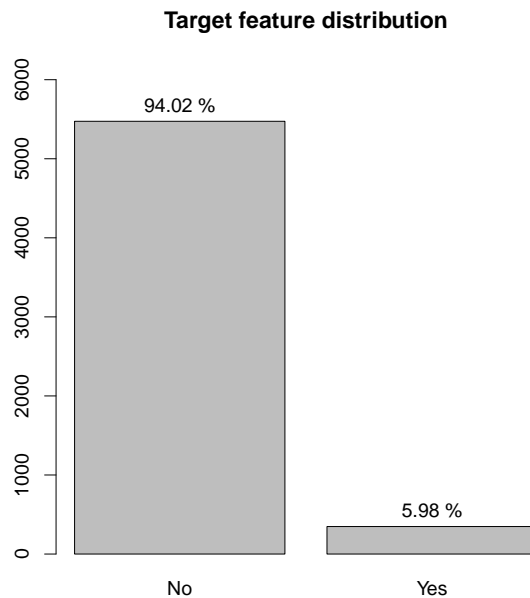
# **Final Homework Report**

NPFL054 Introduction to Machine Learning

**Milan Wikarski**

## Task 1 – Data analysis

Dataset consists of 5822 examples. Each example has 86 attributes. The 86-th attribute *Purchase* with values *Yes* (1) and *No* (0) is the target attribute.



The distribution of the target attribute *Purchase* is heavily skewed towards *No*. The frequency of *Yes* is just 5.98%. This means that the expected precision when randomly selecting 100 examples would be 5.98%.

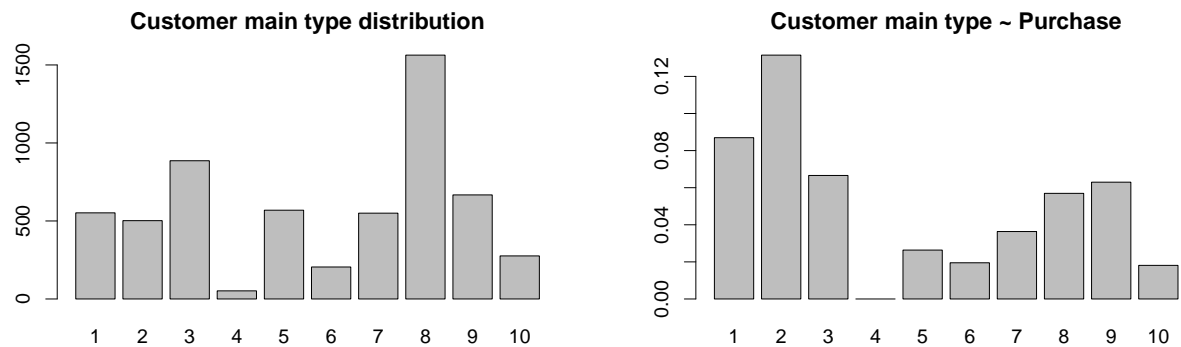
### Task 1a

#### MOSHOODF

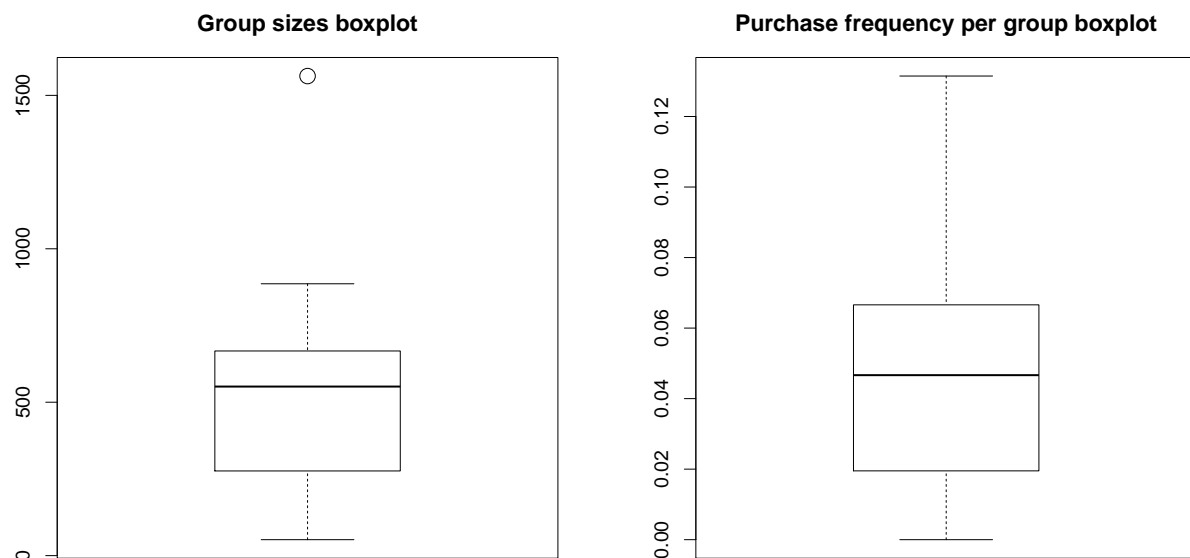
Attribute *MOSHOODF* represents the main customer type. This attribute divides customers into 10 groups described in *L2*. The table below lists the number of customers per group, as well as the percentage of examples in this group with target attribute *Purchase* value *Yes* (1) (ie. the percentage of customers who have purchased the caravan insurance policy):

| ID | Group                 | Size | Purchase Frequency |
|----|-----------------------|------|--------------------|
| 1  | Successfull Hedonists | 552  | 8.7%               |
| 2  | Driven Growers        | 502  | 13.15%             |
| 3  | Average Family        | 886  | 6.66%              |
| 4  | Career Loners         | 52   | 0%                 |
| 5  | Living Well           | 569  | 2.64%              |
| 6  | Cruising Seniors      | 205  | 1.95%              |
| 7  | Retired and Religious | 550  | 3.64%              |
| 8  | Family with Grown ups | 1563 | 5.69%              |
| 9  | Convervative Families | 667  | 6.3%               |
| 10 | Farmers               | 276  | 1.81%              |

This data can also be plotted into two barcharts:



We can further analyze irregularities in groups by plotting the data into two boxplots and looking for outliers:



Group 4 (*Career Loners*) is really small (only 52 members) and contains no positive examples but cannot be classified as outlier. Group 8 *Family with grown ups* is the largest (1563 members), thus can be classified as outlier base on size, but the *Purchase* frequency of this group is slightly below average (5.69%). When it comes to *Purchase* frequency in groups, there are no outliers.

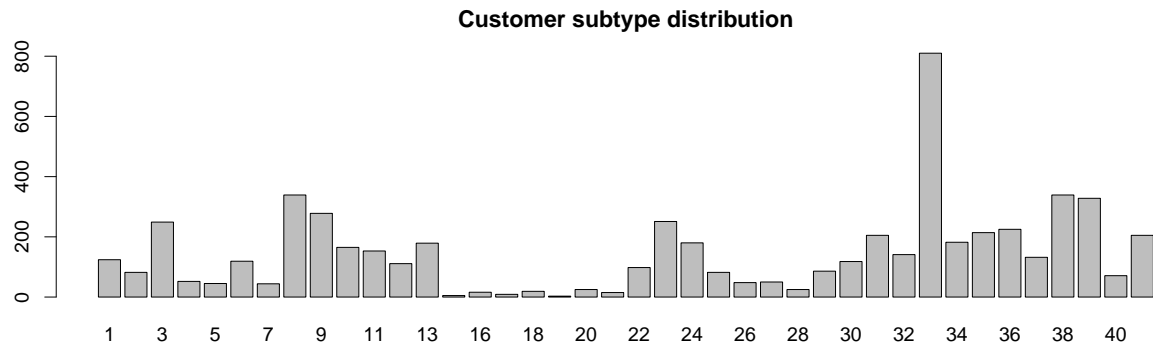
## MOSTYPE

Attribute *MOSTYPE* represents the customer subtype. This attribute divides customers into 41 subgroups described in *L0*. The table below lists the number of customers per subgroup, as well as the percentage of examples in this subgroup with target attribute *Purchase* value *Yes* (1):

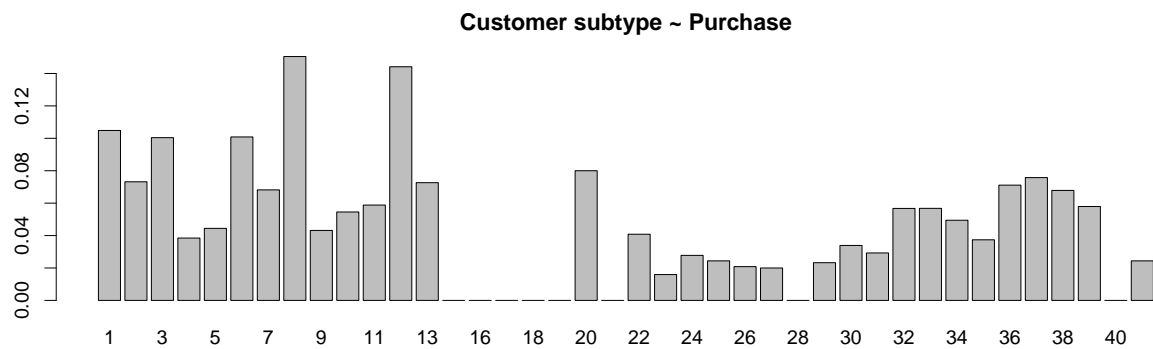
| ID | Subgroup                   | Size | Purchase Frequency |
|----|----------------------------|------|--------------------|
| 1  | High Income                | 124  | 10.48%             |
| 2  | Very Important Provincials | 82   | 7.32%              |
| 3  | High status seniors        | 249  | 10.04%             |

| <b>ID</b> | <b>Subgroup</b>                            | <b>Size</b> | <b>Purchase Frequency</b> |
|-----------|--|-------------|---------------------------|
| 4         | Affluent senior apartments                 | 52          | 3.85%                     |
| 5         | Mixed seniors                              | 45          | 4.44%                     |
| 6         | Career and childcare                       | 119         | 10.08%                    |
| 7         | Dinki's (double income no kids)            | 44          | 6.82%                     |
| 8         | Middle class families                      | 339         | 15.04%                    |
| 9         | Modern                                     | 278         | 4.32%                     |
| 10        | Stable family                              | 165         | 5.45%                     |
| 11        | Family starters                            | 153         | 5.88%                     |
| 12        | Affluent young families                    | 111         | 14.41%                    |
| 13        | Young all american family                  | 179         | 7.26%                     |
| 15        | Senior cosmopolitans                       | 5           | 0%                        |
| 16        | Students in apartments                     | 16          | 0%                        |
| 17        | Fresh masters in the city                  | 9           | 0%                        |
| 18        | Single youth                               | 19          | 0%                        |
| 19        | Suburban youth                             | 3           | 0%                        |
| 20        | Ethnically diverse                         | 25          | 8%                        |
| 21        | Young urban have-nots                      | 15          | 0%                        |
| 22        | Mixed apartment dwellers                   | 98          | 4.08%                     |
| 23        | Young and rising                           | 251         | 1.59%                     |
| 24        | Young                                      | 180         | 2.78%                     |
| 25        | Young seniors in the city                  | 82          | 2.44%                     |
| 26        | Own home elderly                           | 48          | 2.08%                     |
| 27        | Seniors in apartments                      | 50          | 2%                        |
| 28        | Residential elderly                        | 25          | 0%                        |
| 29        | Porchless seniors: no front yard           | 86          | 2.33%                     |
| 30        | Religious elderly singles                  | 118         | 3.39%                     |
| 31        | Low income catholics                       | 205         | 2.93%                     |
| 32        | Mixed seniors                              | 141         | 5.67%                     |
| 33        | Lower class large families                 | 810         | 5.68%                     |
| 34        | Large family                               | 182         | 4.95%                     |
| 35        | Village families                           | 214         | 3.74%                     |
| 36        | Couples with teens 'Married with children' | 225         | 7.11%                     |
| 37        | Mixed small town dwellers                  | 132         | 7.58%                     |
| 38        | Traditional families                       | 339         | 6.78%                     |
| 39        | Large religious families                   | 328         | 5.79%                     |
| 40        | Large family farms                         | 71          | 0%                        |
| 41        | Mixed rurals                               | 205         | 2.44%                     |

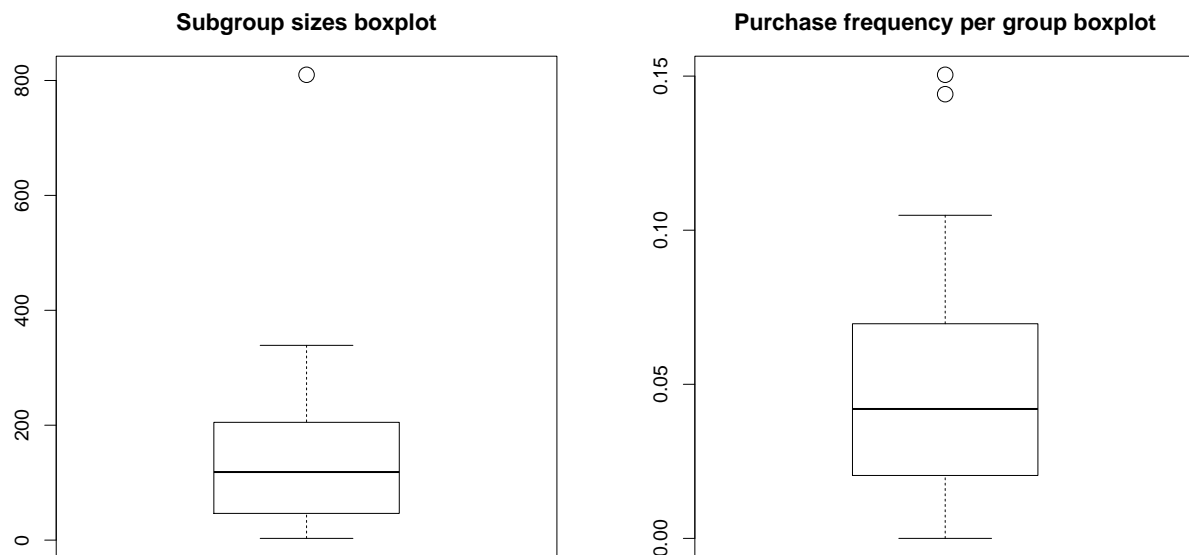
The subgroup sizes can also be visualized using a barchart:



The percentage of people in each subgroup who have purchased the caravan insurance policy can also be plotted to a barchart:



It should be noted that among 5822 examples, there is no representative of subgroup 14 (*Junior cosmopolitan*). To further see, which subgroups might be interesting, we can plot boxplots for size and *Purchase* frequency:



Subgroups 15 – 19 are very small and contain no positive examples but cannot be classified as outliers (the reason for this will be explained in the next chapter). Other subgroups with no positive examples are subgroup 21 *Young urban have-nots*, subgroup 28 *Residential elderly* and subgroup 40 *Large family farms*. Subgroups 21 and 28 are too small to draw any conclusions but there seems

to be a correlation between being a member of subgroup 40 and not purchasing the caravan insurance policy.

The largest subgroup 33 *Lower class large families* is composed of almost 14% of all examples but the *Purchase* frequency in this subgroup is slightly below the average (5.68%). This subgroup is an outlier based on size.

The most promising subgroups are subgroup 12 *Affluent young families* with 111 members and 14.41% *Purchase* frequency, and subgroup 8 *Middle class families* with 339 members and *Purchase* frequency 15.04%, which is the greatest among all subgroups. These two subgroups can be classified as outliers based on *Purchase* frequency.

## Task 1b

After some analysis, it can be seen that *MOSHOOFD* divides the customers into 10 groups and *MOSTYPE* further divides these customers into subgroups. This information can be gained by calling:

```
table(MOSTYPE, MOSHOOFD)
```

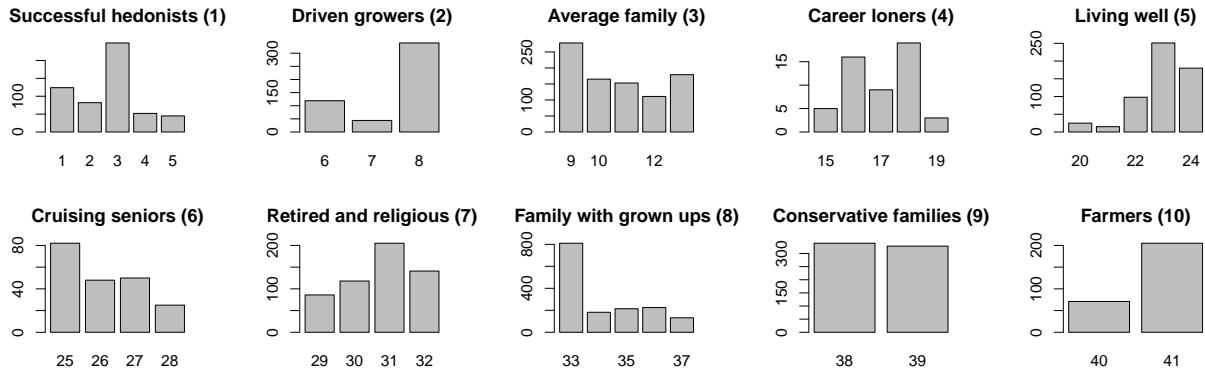
and analyzing the output:

| MOSTYPE | MOSHOOFD |     |     |   |   |   |   |   |   |    |
|---------|----------|-----|-----|---|---|---|---|---|---|----|
|         | 1        | 2   | 3   | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| 1       | 124      | 0   | 0   | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 2       | 82       | 0   | 0   | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 3       | 249      | 0   | 0   | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 4       | 52       | 0   | 0   | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 5       | 45       | 0   | 0   | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 6       | 0        | 119 | 0   | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 7       | 0        | 44  | 0   | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 8       | 0        | 339 | 0   | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 9       | 0        | 0   | 278 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| 10      | 0        | 0   | 165 | 0 | 0 | 0 | 0 | 0 | 0 | 0  |
| ...     |          |     |     |   |   |   |   |   |   |    |

It is clear that all customers of certain subgroup (*MOSTYPE*) belong to just one group (*MOSHOOFD*) and that every customer from each group (*MOSHOOFD*) is assigned one subgroup (*MOSTYPE*).

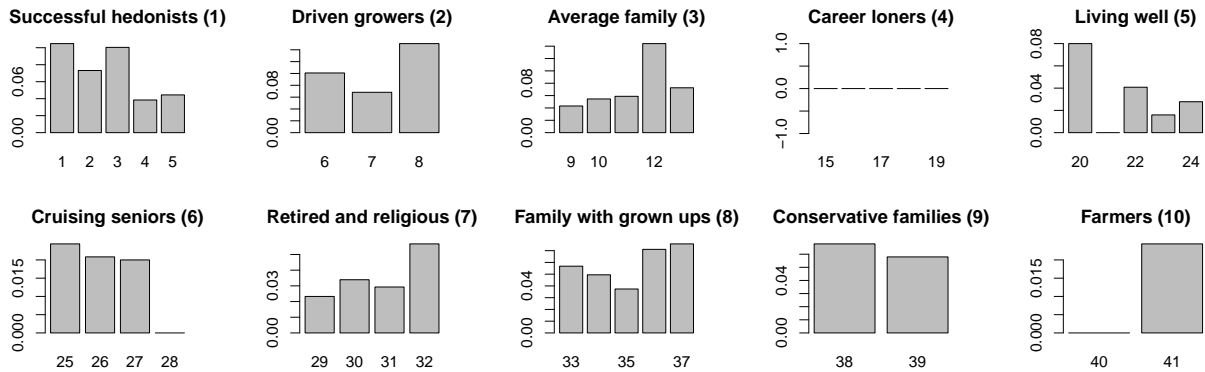
With this information, we can explore the groups (*MOSHOOFD*) in more detail by looking at the size of its subgroups:

### Distribution of subgroups in each of customer main type groups



and the *Purchase* frequency of its subgroups:

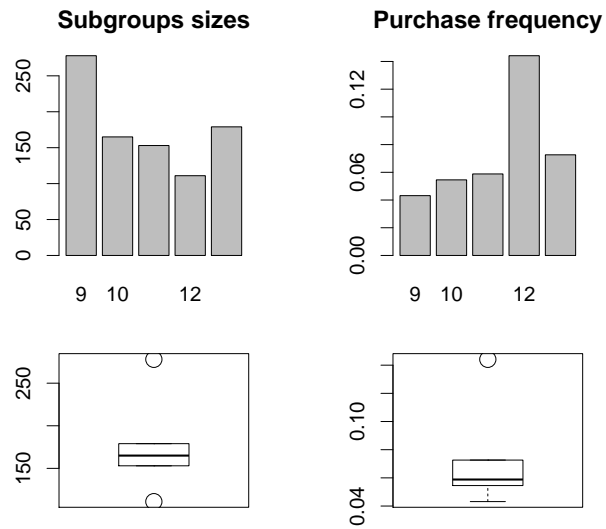
### Purchase frequency of subgroups in each of customer main type groups



This visualization is really useful when explaining why subgroups 15-19 contain no examples. These subgroups are part of group 4 *Career loners*, which as was shown in **Task 1a** contains no positive examples.

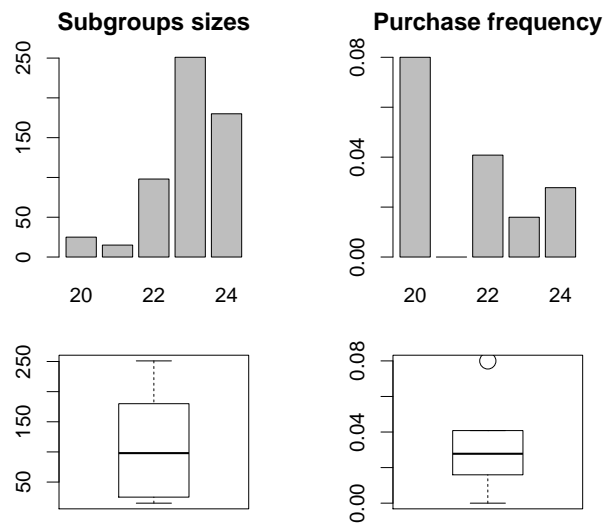
Groups which are interesting are the ones with irregularities in the *Purchase* distribution. Most notably, the likelihood of the members of group 3 *Average family* to purchase the caravan insurance policy is average (6.66%), unless they are part of subgroup 12 *Affluent young families*, in which case it is 14.41%.

### Average family (3)



The same can be said about members of group 5 *Living well* and subgroup 20 *Ethnically diverse*. Although this subgroup can be classified as an outlier compared to other subgroups in the parent group (based on *Purchase frequency*), it contains only 25 members which drastically decreases its importance.

### Living well (5)



Charts like these two (group details) were created for each group and can be seen in `/out/group-<group number>-detail.pdf`.



## **Task 2 – Model fitting, optimization and selection**

## **Task 3 – Model interpretation and feature selection**

## **Task 4 – Final prediction on the blind test set**