Alexandre Picogna & Yasmine Serot

ENSSAT (France)

Data analysis

We will start by observing the general outlines of the histograms. Therefore, we will be studying the continuous features. The age histogram follows a right skewed distribution. Young people are more common than older ones. The sampling weight, while roughly following a uniform distribution before its more frequent value, ends up following what looks like an exponential distribution afterwards. The highest level of education seems to be multi-modal, with more frequency the higher it gets. The capital gains and capital losses look a lot alike. Their minima are the most frequent while their other values are really low. Therefore the gains and the losses are small, which means that people tend to not take risks. The hours per week are inconsistent, with the peak in the middle. This peak is around 40 hours, which is the legal working time in the United States. This is explained by the fact that the majority of the people concerned are from the United States.

We will now look at the missing values. They are only present in the categorical features. We count 6% missing values in the workclass feature, 6% in the occupation feature and 2% in the native-country feature. These percentages are too low so the features concerned should not be removed. Concerning the workclass, either the choices were not extensive enough or this data was not retrieved. If we are in the first case, a study should be done to know what were the actual workclasses. Afterwards, a new category could be made to gather these persons. Concerning the occupation, the percentage is too high to not be a mistake compared to the other values. Moreover, the category 'Other service' should have been enough to gather occupations not falling into any of the other categories. We can guess that there is a problem in the data collect somewhere. Finally, the native-country feature seems to have the same problem, unless countries are not indicated if there are not enough population. If that is the case, we could add a default option.

We observe outliers in three features: age, capital gain and capital loss. While we have seen that there tends to be more young people than old ones, we have a spurt of 90-year-old people. It seems to be out of place and it could be interesting to inquire the reason behind it. It is the same for the two capital features. Their minima are extremely high while the rest is low. Moreover, their other values can be quite spaced out, with their maximum ending up alone. It may be explained by the fact that gains and losses are made by taking risks, and that the bigger the risk, the lesser people taking it. An investigation should still be undertaken. We should also think about the usefulness of this data, since it is so concentrated. It does tell us that there are more gain than loss thanks to the mean, but that gain is less stable than the loss -standard deviation-.

The problems we have seen with the cardinality concerns the marital status. Tough it can get quite complicated, we can wonder about the utility of some denominations. There is also the problem of having a strong cardinality when the percentage of the mode is extremely high: for example with the native country. Finally, the cardinality for the hour per week feature seems to be abnormally high. The maximum of 99 hours should be investigated more closely.