Data Quality Assignment

Colin LEVERGER
Valerian SALIOU
24/05/2016

In this document we perform a data quality review, thus to ensure the dataset we formatted doesn't provide bogus details that may corrupt our further decisions (based on such data).

After generating the datasets + graphs, we proceed to the following analysis:

1. We trim out irrelevant columns (ie: if we don't feel the data will not have an impact in the decision process; we do create a bias to simplify data exploitation)
2. We remove the feature values that are missing (no replacement assumption: e.g. we don't impute)
3. We remove the features with a large miss rate (e.g.: 60%+)
4. We analyze the relevancy of features with an irregular cardinality
5. We analyze the relevancy of outliers (catch invalid outliers: are they unusual entries; e.g.: bogus; or real values, do they give hints on possible data discrepancies?)

Noticeable things about the dataset:

- Race row in the data does not seem to be relevant. We notice a major number of people in the study are from the US, we know that there is about 14% black people in the US; although white people in this dataset are a large majority (without considering Hispanic and black people true ratios): is there a bias on the data?
- The *fnlwgt* looks like a multimodal distribution on histograms, though we don't know what it stands for; we assume we shall nuke it?
- In this dataset for the continuous data, there is no cardinality <10 (this is a very simple observation)
- In the *capital-gain*, there is a lot of "0" s (approximately 29,000). This seems to be strange at a first glance, and it should be interesting to investigate in order to take a decision. Same for capital losses. On business matters, we shall ask ourselves whether those zeros are normal (expectable), or abnormal.
- There is a lot of "private" in *workclass*. Is it relevant to keep them and why is there so much of it?
- For the *education* row, we can observe a normal sleek rate.
- We should put in relation age data & never married data? Since we have a lot of people who never married, it could be interesting to put in relation the *age* and *never_married* fields because they are surely linked (in a certain way).
- *occupation*: we have a lot of "?" s (count: 1742); this involves a bias since we have to trim out some features.
- In the *native_country* there is more than 95% of people from the US. Since we are working for an insurance company that should be located in the US, we assume that this column may not be useful/relevant.