Data Quality Issues

Student name: Maximilian Mihoc Number: C12728559 Date: 15-01-2016

Missing Values

The percentage of missing values in the tables is not bad at all. The only features that have some missing values are Work class, Occupation and native country. The % of missing values for Work class feature is 5.6% witch is not that big so it would be extreme to eliminate this feature. Same thing applies for the Occupation feature which has a missing value percentage of the same value as the previous feature, 5.6%. Taking into account that this feature are both related to people jobs and the same percentage of 5.6 is missing for both features, I am thinking that the data of both these features was collected in the same time and from the same people. Maybe they were not confident with giving this kind of information and this can be investigated with the business. The last feature that has some missing values is Native-Country. This feature has 1.8% of missing values and it would be again extreme to remove it. This is something that can be investigated with the business to see what the reason for this missing values is. It may be the case that people were not confident in giving nationality details.

Cardinality

The cardinality of the features seems good for categorical features but the continuous features have some irregularities. Taking into account that there are 30940 instances, most of the continuous features have less than 100 in the cardinality which seems unusual. Some features have a good distribution as it covered almost all possibilities that the feature value could have. If I take as example the AGE feature, having the minimum value of 17 and the maximum value of 90, maximum cardinality could be 73 and in our table it is 72 which means that only one is missing. The HOURS PER WEEK feature also have a small cardinality because this features can be very similar for many people and I don't think there is a problem with it. More than that, I think this feature could have smaller cardinality as the maximum of 99 hours per week seems a bit too much and needs more investigation. The FNLWGT feature has a unique value for almost every instance which is good for a continuous feature. CAPITAL GAIN and CAPITAL LOSS features have very small cardinality in my opinion and more investigation needs to be made here.

Outliers

First thing to be observed in the continuous feature table is about the CAPITAL_GAIN and CAPITAL_LOSS both having a lot of zero values. More investigation needs to be made here to see if the data is valid or not. Looking at the 1st quartile, median and 3rd quartile, which are all 0, we can say that more than 75% of instances have 0 for these features. A value of 0 here means that more than 75% of people do not have any capital gain or loss which is not impossible but I still think that more examination is needed. The maximum value of 99,999 as maximum for the CAPITAL GAIN seems very big and investigation with the business is required here. The instances that have this value need to be found in the dataset and they need to be investigated with the business.

A big difference in the FNLWGT feature between maximum value and the 3rd quartile seems unusual and in my opinion, the instances with the big maximum values should be localised and exanimated. The deviation from the norm is very big here and it may be the case that different instances were placed together.

An observation in the cardinality table, which is not necessarily an issue is that the Education and Education-num features are identical, the only exception being the representation of data, string vs number, and I think that we can get rid of one of those two features.