Machine Learning 1: CA1 Report by Jonathan Noble(C15487922)

Categorical vs Continuous Data Type

First of all, I think it's important to note that ID is removed since it is not a target feature anyway. I also normalized one of the categorical features by dropping the education_num since it has the equal value as education. Having said all that, I used the python feature select_dtypes which separated "int64" data types (which exhibited as the continuous features) apart from "object" data types (which exhibited as the categorical features). Bear in mind that using select_dtypes does not equate to an automatic key to separating features from categorical and continuous, I have assessed the features meticulously beforehand by knowing the difference of each feature between categorical and continuous (categorical contain a finite number of categories or distinct groups whereas continuous are numeric variables that have an infinite number of values between any two values) and ensured that each feature corresponds to their own dtypes… and clearly enough, they did.

Identifying Data Quality Issues & Handling the Issues

I believe the second priority of data exploration is to identify any data quality issues. It is defined as anything unusual about the data. In the data of CA1's case, I have managed to outline the missing values The following points below are the ones that normally occur:

Missing values

Analysing the data quality report, we can see that only the table of categorical features have the missing values from: workclass (5.94%), occupation(5.97%) and native-country(1.82%). After having another inspection, it is obvious that the correlation between these three features is a result of having high cardinality values. They are among the top three highest cardinal values in the table.
A complete case analysis can be done for the features workclass and occupation to remove instances that are missing the value of the target feature. Given the fact that native-country has the highest cardinality of all in addition to having the lowest missing value, an imputation can be a good option to accommodate the problem for that specific feature.

Irregular Cardinalities

There are three main points to determine Irregular Cardinalities in the data:

- Features with a cardinality of 1
- Features incorrectly labeled
- Categorical features having higher cardinalities than expected

Starting from continuous features, we can see that fnlwgt has the highest cardinality among the rest and its value is close to the number of instances in the dataset - this normally happens to a continuous feature and it is fine. Additionally, it is also expected that capital-gain and capital-loss may also have high cardinalities but this is due to many instances having a null value. Jumping towards categorical features, none of them had violated any of the points above and seems okay as well. For

instance, Sex have the cardinality of 2, target have 2 whereas native-country of 49 values, which appears to have the highest among the features, can be said that it is still reasonable enough since there are 195 countries in the world.
With that said, the quality of data of both continuous and categorical features are on the right track and should be left the way it is.

Outliers

Outliers are values that lie far away from the central tendency of a continuous feature. It is defined to have two kinds of outliers: valid outliers and invalid outliers. There are also two approaches to identify outliers within the dataset:

- Examining the minimum and maximum values for each feature
- Comparing the gaps between the median, minimum, max, 1st quartile and 3rd quartile values. If the gap between the 3rd quartile and the max value is larger than the gap between the median and the 3rd quartile, this could mean that the max value is likely to be an outlier. Concurrently, if the gap between the 1st quartile and the minimum value is larger than the gap between the median and the 1st quartile, this suggests that the minimum value is likely to be an outlier.

First approach is that no negative values are shown in any of the features so it is assumed that they are all plausible values. By proceeding towards the second approach, all of the features apart from hours-per-week appears to have unusual maximum values and suggests that they are outliers. Similarly, the feature age has an unusual minimum value and also suggests that it is an outlier. It is worth note taking that with the second approach, it is likely that those outliers found are valid outliers, so they are a data quality issue due to valid data.
We can handle outliers by using the clamp transformation. This clamps all values above an upper threshold and below a lower threshold which would remove the offending outliers.