Data Quality Report

Erifeoluwa Jamgbadi

12/10/2021 – 31/10/2021

—

Machine Learning for Data Analytics

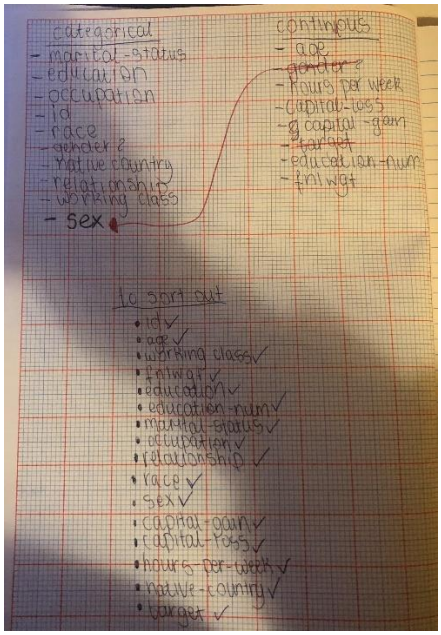
—

C18387973

Table of Contents

Features (categorical vs. continuous)

I sorted out which one was categorical and which one was continuous on paper, this helped me to identify, which data belonged to either of the two.



Continuous Data from csv

	A	B	C	D	E	F	G	H	I	J	K
1	FEATURE	Count	% Miss	Card.	Min	1st Qrt.	Mean	Median	3rd Qrt.	Max	Std Dev.
2	age	30940	0	72	17	28	38.56076	37	48	90	13.6394
3	fnlwgt	30940	0	20880	12285	117849	189786.4	178384	237318	1484705	105406.4
4	educatio	30940	0	16	1	9	10.08125	10	12	16	2.569967
5	capital-	30940	0	119	0	0	1081.813	0	0	99999	7443.773
6	capital-	30940	0	91	0	0	86.56997	0	0	4356	401.706
7	hours-	30940	0	93	1	40	40.40892	40	45	99	12.33694

Categorical Data from csv

	A	B	C	D	E	F	G	H	I	J
1	FEATURE	Count	% Miss	Card	Mode	Mode Fre	Mode %	2nd Mode	2nd Mode	2nd Mode %
2	id	30940	0	30940	tr26861	1	0.003232	tr16342	1	0.003232
3	workclas	30940	0	10	Private	21575	69.73174	Self-emp	2406	7.776341
4	educatio	30940	0	16	HS-grad	9976	32.24305	Some-col	6938	22.42405
5	marital-	30940	0	7	Married-c	14201	45.89851	Never-ma	10167	32.86037
6	occupati	30940	0	15	Prof-spec	3932	12.70847	Craft-rep	3887	12.56303
7	relations	30940	0	6	Husband	12496	40.38785	Not-in-fa	7904	25.54622
8	race	30940	0	5	White	26442	85.46218	Black	2965	9.583064
9	sex	30940	0	2	Male	20705	66.91984	Female	10235	33.08016
10	native-	30940	0	42	United-St	27719	89.58953	Mexico	607	1.961862
11	target	30940	0	2	<=50K	23506	75.97285	>50K	7434	24.02715

Potential data quality issues with the data

- **Missing values**

To find the missing data, I filtered it out in excel:
Id – no missing values
Age – no missing values
Workclass – has missing values
Fnlwgt – no missing values
Education – no missing values
education-num – no missing number
marital-status – no missing values
occupation – has missing values
relationship – no missing values
race – no missing value
sex – no missing value
capital-gain – no missing value
capital-loss – no missing value
hours-per-week – no missing value
native-country – has missing value
target – no missing value

Analysis:

- There is a correlation between working class and occupation. People who have both the data from working class and occupation missing could mean has never worked because from the data, there are some who have occupation missing, but have working class and its says never worked. I noticed that these people were in some sort or school or still 18 years old. It also depends on age and education because it shows where both these are blank, the people are just 18, still in some form of school e.g. high school or have probably retired and/or couldn't find a job/never worked, so I would either remove them or better yet, just put them as never worked. I would more likely put their working class as never worked, if they were 18 or in some sort of education e.g college. If they were of retiring age, I would fill those missing data of working class with retired.

- Those that have a missing value of native-country has a more chance of having a target of less/equal to 50000k, than with more than 50000k. There could be a number of reasons for this, which the data would not allow me to know, but I would be able to make different guesses.

- **Outliers**

To find the missing data for the continuous data, I worked it out in excel. For each feature heading of the continuous data e.g education num, capital-gain etc, I found the outliers with this method:

- Got the 1st quartile
- Got the 3rd quartile
- Got the interquartile range
- Used those values to get the upper bound
- Used those values to get the lower bound
- Then the values of those features were compared

Analysis:

- Capital-gain and capital-loss a high maximum value compared, especially compared to the median and 3rd quartile.
- For my analysis, I found out, a person would be more likely to get a capital gain, if they were a male, lived in the United States and their person target is more than 50000, then they would be more likely to get a capital-gain that is higher.
- For others, their capital-gain would be smaller if their target is less or equal to 50000.
- This analysis makes sense. I believe from analyzing the data, that capital-gain is dependent on target.

- **Feature cardinality issue**

Analysis:

- For the education-num, there is nothing wrong with this data, but I don't think its very useful for predictions, especially since the education is there.

My opinion as to what should be done to address these quality issues

I also put some of my opinion on what should be done to address some of the data quality issues in my analysis in the potential data quality section.

I believe imputation should be used to replace the missing value feature with a plausible estimated value based on the feature values that are present. For the continuous features, I would replace the missing values with either the mean or median and for the categorical features, I would replace it with the mode. The median, mean or mode are good options to use because they represent the data and what the missing value could be or how they might be within that range, so that's the main reason I would use them to replace missing values. I would only use imputation, if the missing values are less than 30% because then the data wouldn't be as accurate or won't be good quality especially, if the missing data replaced is definitely over 50%.

I would use clamp transformation for the outliers. This will help to make the data quality better because I would be removing the offending outliers. The outliers that have the maximum value 99999(capital-gain) are too high, especially against the min 0 (capital-gain), so I believe it would be better to remove them. Therefore, for capital-gain, lower limit of 0 and upper limit of 99999.

For the cardinality issue, I would remove the education-num column/feature because I don't think it's very useful especially since we have the name of the education in the data.