Data Quality Homework : Evaluation of Data Quality Metrics

Trishita Patra - MDS202440

The goal of this homework is to evaluate Data Quality Metrics using the following and and derive the Data quality metrics.

1. Ydata Profiling
2. Amazon PyDeequ software
3. Great Expectations

1. Data Profiling

Provides an automated, detailed EDA (Exploratory Data Analysis) report on a dataset.

- Descriptive stats: mean, std, min/max, quantiles, etc.
- Missing values: percent missing per column
- Correlations: Between attributes/columns.
- Duplicate rows
- Column type inference (categorical, numerical, etc.)

The PDF report for this can be found attached at the end. The html report is also attached with the zip file.

2. PyDeequ

A Scala-based library from Amazon (wrapped in PySpark) that checks for data completeness, distinctness, entropy, and compliance to various conditions on large datasets using Spark.

Dataset Overview

- **Dataset Size:** 48,842 rows

Completeness

Passed Constraints (Success):

- The `age` column has no missing values.
- The `education` column has no missing values.
- The `marital-status` column has no missing values.
- The `relationship` column has no missing values.
- The `race` column has no missing values.
- The `sex` column has no missing values.
- The `income` column has no missing values.
- All values in the `age` column are non-negative.
- All values in the `income` column are within the expected set: `<=50K`, `>50K`.
- All values in the `sex` column are either `Male` or `Female`.

Failed Constraints (Failure):

- `workclass` column is only approximately 94.27% complete.
  - About 5.73% of the rows have missing (`null`) values.
- `occupation` column is only approximately 94.25% complete.
  - About 5.75% of the rows have missing values.
- `native-country` column is only approximately 98.25% complete.
  - About 1.75% of the rows have missing values.

Approximate Distinct Value Counts per Categorical Column

Column	Approx. Distinct Values
workclass	8
education	15
race	5
native-country	39
sex	2
relationship	6
occupation	14
income	2
marital-status	7

Summary of Column-Level Metrics

Mean Values

Column	Mean Value
age	38.64
education-num	10.08
hours-per-week	40.42
capital-gain	1079.07
capital-loss	87.50

Correlation Between Columns

Columns	Correlation
age, hours-per-week	0.072
education-num, capital-gain	0.125

Entropy (Measure of Categorical Diversity)

Column	Entropy Value
education	2.03
race	0.55

Distinctness

Column	Distinctness Value
occupation	0.0003041

Compliance (Proportion of rows satisfying a condition)

Column / Rule	Compliance
Income : >50K	0.2393
Native-country : India	0.0031
Race : Amer-Indian-Eskimo	0.0096
Relationship : Not-in-family	0.2576
Occupation : Armed-Forces	0.0003
Marital-status : Never-married	0.3300
Education : Bachelors	0.1643
Workclass : Without-pay	0.00043

The PyDeequ checks closely mirror the findings from YData Profiling.

Great Expectations

A framework for writing unit-test-like expectations for your data, which can be reused across workflows or pipelines.

Data Quality Report: Adult Dataset

Numerical Columns | Range Checks

We assessed the range of values for all numerical columns to ensure they fall within expected limits:

- **age**: 17 to 90
- **fnlwgt**: 12,285 to 1,490,400
- **education-num**: 1 to 16
- **capital-gain**: 0 to 99,999
- **capital-loss**: 0 to 4,356
- **hours-per-week**: 1 to 99

Categorical Columns | Expected Value Sets

We validated that each categorical column only contains values from known and accepted categories:

- **sex**: Male, Female
- **race**: White, Black, Asian-Pac-Islander, Amer-Indian-Eskimo, Other
- **income**: <=50K, >50K, <=50K., >50K. (including trailing dot variations)
- **workclass**: Includes Private, Self-emp, Gov, Without-pay, Never-worked, etc.
- **education**: Includes Bachelors, HS-grad, Masters, 10th, Preschool, etc.
- **marital-status**: Married, Divorced, Never-married, etc.
- **occupation**: Includes Sales, Tech-support, Armed-Forces, etc.
- **relationship**: Wife, Husband, Not-in-family, etc.

Null Checks

We confirmed the presence of non-null values in most key columns:

- Passed: age, education-num, income
- Failed: workclass, occupation, native-country

The failed columns contain missing values as expected based on the data profiling summary. These failures confirm the earlier observations.

Median and Data Types

We checked that:

- The **median** of both `capital-gain` and `capital-loss` is zero, consistent with their skewed distribution.
- The data types of `capital-gain` and other numerical columns are valid (int or float).

Target Column Distribution

We evaluated class imbalance in the **income** column using KL divergence:

- Expected distribution: 75% <=50K and 25% >50K
- The divergence was within threshold, indicating no severe imbalance.

Validation Summary

- **Total Expectations Evaluated**: 24

- **Success Rate**: 87.5%
- **Failed Expectations**: 3

Failed Expectations

1. **workclass**: Contains missing values
2. **occupation**: Contains missing values
3. **native-country**: Contains missing values

These failures are expected and consistent with earlier data profiling using YData Profiling, which flagged these columns for null values.