Data Quality, Privacy and Ethical AI

Final take home project

Roudranil Das
MDS202227
roudranil@cmi.ac.in

Contents

1 Objective

Implement differential privacy on broadband usage Data

2 Exploratory data analysis

2.1 Dataset information

According the data documentation, both datasets contain the following columns:

- `ST`: is the 2 letter abbreviation of states in the United States
- `COUNTY ID`: 4 to 5 digit code used to represent the county (last 3 digits) and the state (first digit or first 2 digits)
- `BROADBAND AVAILABILITY PER FCC`: percent of people per county with access to fixed terrestrial broadband at speeds of 25 Mbps/3 Mbps
- `BROADBAND USAGE`: percent of people per county that use the internet at broadband speeds based on the methodology explained above

2.2 Dataset cleaning

First of all, we replace all the spaces in the column names with underscores, in order to make it easy to write SQL queries with the column names.

Following this, we observe that the `BROADBAND AVAILABILITY PER FCC` and `BROADBAND USAGE` columns are of string datatype, where they should be float. This indicates that there are missing or anomalous data. Upon closer examination, we found that the missing data was encoded as `' - '`. This was replaced by None. Any other form of imputation would be unsound, as broadband usage depends on many local factors. For the rest, the trailing and leading spaces were removed, and the data type was changed to float.

This was done for both the datasets and for both the columns.

2.3 Creation of the dataset metadata

In order to properly use smartnoise smartnoise, we had to create a yaml file with the table metadata. The metadata contains information about the table name, the column names, the column data types, lower and upper bound of the balues in the columns (if applicable) and if row privacy is required. Row privacy was required for both of the datasets.

3 Comparison of results after applying differential privacy

Here, we try to find the average broadband usage for each state. To do that we group by the state column and take the average of the broadband usage column. The MAE and RMSE between the differentially private results and the original results are given below.

Dataset	ϵ	Mean Absolute Error	Root Mean Squared Error
November2019	**0.1**	0.4783	1.6577
	0.2	0.1746	0.4390
October2020	**0.1**	0.2171	0.4130
	0.2	0.2298	0.6386

Table 1: Comparison of results for differential privacy

One point to note is that the differentially private results sometimes contain a negative value for broadband usage.

We did not use the zipcode level dataset, as the data there is already made differentially private.