CS211 Final Project Report

Isabelle Gagnon
December 10th, 2021

Problem

Information regarding loans can be sensitive, and most people do not want this information accessible to the public. For this project, I used a dataset from Kaggle.com, which is easily accessible online. The loan dataset I used for this project assigns a unique loan ID to each individual. Other information includes, loan status (whether the loan was paid off on time, sent to collection, or sent to collections and paid off), principal loan amount, terms (the payoff schedule), the origination date of the loan, the due date, pay off time (the date in which the loan was paid off), and days past due (the amount of days the loan was past due). Age, education level, and gender data is also included, which is potentially identifying information (PII), and could be linked back to specific individuals.

Description

To accomplish the results of this project, I used the above threshold method, the Laplace mechanism twice, and the sparse vector technique to automatically determine a clipping parameter. The Laplace mech is used each time with 1/3 of the privacy budget. This allows for epsilon to be satisfied by differential privacy, and there is a wide range of possible values for b. I was then able to derive a differentially private average of the data by determining the noisy sum and the noisy count using the Laplace mechanism, and then dividing the noisy sum by the noisy count.

I used the above strategy to determine the average of different columns in the dataset. The actual average was calculated by dividing the sum of the data by the length, and the differentially private data was determined by the above explanation. I was also able to compute the majority of occurrences in the datasets that contained non-numerical data. This was calculated by using the score, and report noisy max methods from the text. Report noisy max is able to satisfy epsilon differential privacy because it releases only the identity of the element with the largest noisy count. Noise is added to the score, the index of the maximum score is found, and the element with that index is returned.

In order for the non-numerical data to be used, I used a uniqueness method to determine the different values within the specific columns. For example, the gender column contained male and female, and the education column contained different levels of education -- high school or below, college, bachelors, and masters or above. These were unique values within the columns. I was also able to split the education data into two levels, college educated and not college educated. This was helpful in narrowing down the levels and comparing data.

Results

The results of this project were interesting. I calculated the average age and average principal amount, and the average principal amount of those 25 and under, and those over 25. With these averages, the differentially private (DP) averages were also calculated. The average age was 31, and the DP age was 33, with a percent error of 6.4%. The average principal loan amount was $943, and the DP amount was $953, with a percent error of 1.1%. The average principal amount of those 25 and under was $943, and the DP amount was $941, with a percent error of 0.1%. The average principal amount of those over 25 was $942, and the DP amount was $12, with a percent error of 98.6%. The last percent error was surprising, but it could be an error or outlier. The rest of the DP averages were fairly close to the actual average, meaning the addition of this privacy could be accurate enough to be beneficial.

For the maximum occurrence counts on the non-numerical data, I calculated gender, education level, loan status, maximum occurrence of education level of those with past due loans, and which gender had maximum occurrences of past due loans. The results showed more males took out loans than females, college level individuals took out the most loans, a majority of people paid off their loans on time, college level individuals have the most occurrences of not paying off their loans on time and letting them go to collections, and males are more likely to not pay off their loans than females. All of this information matched the differentially private results except for one -- the DP result indicated those with high school education level or below have the most occurrences of not paying off their loans on time. Taking these results into account, the addition of this privacy could be accurate enough to be beneficial.