FL : A Flexible Federated Learning Package for Real-World and Simulated Applications

Justice Akuoko-Frimpong, Michael Kaye, and Jonathan Ta

December 2023

Introduction

For insurance companies and customers, it is important to understand what factors contribute to any medical bill amounts for hospital stays. This information would help customers budget for planned or emergency treatments. It would also help insurance companies make decisions about what to cover and what premiums are necessary to profit from covering certain medications, treatments, etc. With this motivation in mind, the goal of our proposed study would be to make inferences about the factors associated with billing amounts.

Some studies have already examined influences on hospital billings in more specific contexts. Gender and type of care have been linked to daily intensive care unit costs (Dasta et al., 2005). In addition, a study by Gams et al. (2017) demonstrated associations between severe odontogenic infection hospitalization costs and age, use of antibiotics, and diabetes status. Length of stay was included, as well, because of how variable it can be in the context of general hospital admissions, as opposed to looking at admissions for a specific condition or treatment. An interaction term was also added between length of stay and medical condition to test if the relationship between length of stay and cost was influenced by the medical condition.

According to the American Hospital Association (2023), there were over 34 million hospital admissions in the US between January and May of 2023. With such an abundance of data, it might seem that answering these questions could be done by obtaining hospital admissions and billing data from the insurance companies. However, insurance companies must protect patient privacy and adhere to the Health Insurance Portability and Accountability Act of 1996, also known as HIPAA.

Federated learning solves the issue of protecting patient data while also allowing insurance companies to share data with us for our analysis. We created an R package and Shiny app to perform a linear regression using federated learning. With our Shiny app, insurance companies can produce the necessary summary statistics from their hospital admissions data. The functions in our package can then generate the same regression results that would be obtained using the patient-level data, using only the summary statistics.

Methods

Federated learning math theory

Federated Learning is an emerging machine learning paradigm that aims to train a collaborative model while keeping all the training data localized (Yang et al., 2022). Data communication from client devices to global servers is not necessary. Rather, the model is trained locally using the raw data on edge devices, hence improving data privacy. The local modifications are aggregated to build the final model in a shared way.

In clinical settings, to maintain patient data security, federated learning still has to be implemented carefully. However, it has the potential to tackle some of the challenges we faced by approaches that require the pooling

of sensitive clinical data. Clinical data can be collected inside an institution's security safeguards for federated learning. Each individual maintains ownership of their own clinical data. Federated learning allows teams to create bigger, more varied datasets for algorithm training, even as it becomes more difficult to retrieve sensitive patient data as a result. By using a federated learning strategy, various healthcare facilities, research institutes, and hospitals are also encouraged to work together to create a model that might be advantageous to all of them.Here are some reasons why federated learning matters (Shastri, 2023):

Privacy: Federated learning allows training to happen locally, avoiding potential data breaches, in contrast to traditional approaches that send data to a central server for training.

Data security: is ensured by sharing only summary statistics updates with the central server.

Access to heterogeneous data: is ensured via federated learning, which makes data dispersed across many companies, places, and devices accessible. It allows for secure and private training of models on sensitive data, like financial or medical data. Additionally, models may be made more generalizable with increased data diversity.

How does federated learning work?

At the central server is a general baseline model. The client devices receive copies of this model, and they use the local data they produce to train the models. Individual models improve with time, becoming more tailored to the user's needs. In the next phase, secure aggregation techniques are used to exchange the updates (summary statistics) from the locally trained models with the central server's primary model. This model creates new learnings by averaging and combining various inputs. Once the model in the central server has been re-trained with the new summary statistics, it is shared with the client devices again for the next iteration. The network diagram in the appendix illustrates the flow of federated learning approaches.

Fitting a linear regression model

Suppose we have k groups in a study. Each group has different data but identical columns. To answer a research question with a linear regression model, assume independence of the responses from individuals across the k groups. The coefficients for the regression can computed using the following user-specific summary statistics SSX, SSY, $SSXY$, and n from all k groups to obtain the estimates of the coefficients using the formula:

$$\hat{\beta} = (\sum_{k=1}^{K} X_k^T X_k)^{-1} \times (\sum_{k=1}^{K} X_k^T y_k)$$

where X_k is the design matrix and Y_k is the response vector for the k-th group.To further evaluate the significance of the coefficients, standard errors of the estimated coefficients will be calculated using the formula below, which is also in the federated learning form:

$$\widehat{se}(\widehat{\beta}_j) = \sqrt{\widehat{\sigma}^2 (X^\top X)^{-1}_{jj}}, \text{ with } \widehat{\sigma}^2 = \frac{\widehat{\epsilon}^\top \widehat{\epsilon}}{n-p},$$

where

$$\widehat{\epsilon}^\top \widehat{\epsilon} = \sum_{k=1}^{K} y_k^\top y_k - 2\widehat{\beta} \sum_{k=1}^{K} X_k^\top y_k + \widehat{\beta} \left(\sum_{k=1}^{K} X_k^\top X_k \right) \widehat{\beta}$$

P-values and t-statistics can be computed using these estimated values. After the calculation of all the coefficients and corresponding variances, we conduct a t-test to test the significance of each coefficient.

FL Package

We created a federated learning package "FL" that is hosted on github. This package facilitates the federated learning analysis in a modular format with 3 key functions.

The first function ("FL_local_summary") takes in a csv input which includes the specified data set as well as the desired regression formula. It then runs the local server analysis with a clean output of all the summary statistics. The local servers will then send that summary output to the central server with no additional work. The package will facilitate this through the use of an R Shiny app (to be discussed in detail later).

The second function ("FL_combine") is designed to be used at the central server level and combines all the summary statistics from the local servers as described earlier in the report. The input is the exact output from "FL_local_summary".

The third function ("FL") runs the federated learning analysis to produce the results of the study. The input for this function is the exact output from "the second function"FL_combine".

This seamless integration of inputs and outputs allows the FL package to run a complete federated learning analysis.

Dataset

Table 1 below shows summary statistics for all the variables in the Aetna data set, as an example of how the local data is structured. All of the data used were simulated to appear in a manner that would be straightforward for insurance companies to provide.

Table 1: Descriptive statistics for Aetna dataset

	Overall
	(N=2025)
Age	
Mean (SD)	51.9 (19.8)
Median [Min, Max]	52.0 [18.0, 85.0]
Gender	
Female	1012 (50.0%)
Male	1013 (50.0%)
Medical Condition	
Arthritis	308 (15.2%)
Asthma	368 (18.2%)
Cancer	362 (17.9%)
Diabetes	324 (16.0%)
Hypertension	348 (17.2%)
Obesity	315 (15.6%)
Billing Amount	
Mean (SD)	25800 (14100)
Median [Min, Max]	25800 [1010, 50000]
Admission Type	
Elective	659 (32.5%)
Emergency	706 (34.9%)
Urgent	660 (32.6%)
Medication	
Aspirin	412 (20.3%)
Ibuprofen	402 (19.9%)
Lipitor	440 (21.7%)
Paracetamol	368 (18.2%)
Penicillin	403 (19.9%)
length.of.stay	
Mean (SD)	16.6 (8.63)
Median [Min, Max]	17.0 [2.00, 31.0]

Although there are a limited number of groups for the categorical variables in this simulated data, all of the functions in FL can accommodate any number of unique categories. To combine the local summary statistics, each insurance company must utilize the same categories for each variable. Collaboration with the various insurance companies to standardize how these categorical variables are reported would take place before data gathering begins.

Application and Results

R Shiny

The Shiny app allows insurance companies to generate all of the necessary summary statistics for the federated learning linear regression by simply uploading their data set as a csv file. Table 2 below shows all the variables and their respective types that must be contained in the csv file.

Table 2: Required CSV column headers and data types

Variables	Type
Age	double
Gender	character
Medical Condition	character
Billing Amount	double
Admission Type	character
Medication	character
length.of.stay	double

After the csv has been uploaded, 4 different sets of summary statistics will be generated to allow for flexibility in how the final model is constructed. The different options considered were whether to include an age-squared term (Age.Sq) and/or an interaction between Medical Condition and length.of.stay (Int). By obtaining all 4 sets of summary statistics from each company, we can later perform the hypothesis tests necessary to decide if age-squared term and the interactions should be included in the final model. The app presents the summary statistics for each potential model in separate tabs.

Validation

Our project treated different insurance companies as the local servers and used the equation [***]. For all five local servers we ran the "FL_local_summary" function and sent only the summary statistics to the central server. We then ran the "FL_combine" function. Finally, we ran the "FL" function and got the results below.

To prove that our federated learning process was a success we ran the well established lm() function on the oracle (combined) dataset and compared the results to our federated learning process. As can be seen below, we have exact matches in the desired statistic. Full linear regression results using both methods are shown in Table 3 in the Appendix.

```
all.equal(unname(coef(oracle)), FL$Estimate)
```

```
## [1] TRUE
```

```
all.equal(unname(oraclesum$coefficients[, "Std. Error"]), FL$Std..Error)
```

```
## [1] TRUE
```

```
all.equal(unname(oraclesum$coefficients[, "t value"]), FL$t.value)
```

```
## [1] TRUE
```

```
all.equal(unname(oraclesum$coefficients[, "Pr(>|t|)"]), FL$p.value)
```

```
## [1] TRUE
```

Discussion and Limitations

Our package and federated learning in general has some limitations. Model diagnostics are difficult to carry out on the scale of complete data, but they may be done at the individual user, group, or site level utilizing subsets of data which will be a little difficult to generalize to the whole model in the central server. Verifying the quality of data and locating anomalies or influential data instances is difficult. Since more observations increase the variability of the data distribution, an outlier in a local data set may not always represent an outlier in the entire set. The flexibility of data analysis operations at individual users'/sites is limited since all users/sites are needed to concur on gathering the same data and doing the same analysis in order to provide the necessary statistics. Additionally, a federated learning analysis requires trust in the local level statisticians as their analysis up to the summary statistics cannot be seen by the central server.

Our package mitigates the risk of statistician error by only requiring local servers to format data into the input csv format. Data organization has far lower risk than data analysis which is done completely by the FL package. Additionally, incorrect formatting will lead to an error message providing the local servers an opportunity to correct any mistakes made. Out package is also scalable, in that we can easily add additional local servers or update the desired regression formula.

By using federated learning via the FL package researchers are able to collect far more data than otherwise available as data privacy issues are controlled through the federated learning technique.

References

- American Hospital Association (2023, May). *Fast facts on U.S. Hospitals.* (https://www.aha.org/statistics/fast-facts-us-hospitals)
- Dasta, J. F., McLaughlin, T. P., Mody, S. H., & Piech, C. T. (2005, June). *Daily cost of an intensive care unit day: The contribution of mechanical ventilation.* [Critical Care Medicine, 33(6), 1266-1271]. DOI: 10.1097/01.CCM.0000164543.14619.00.
- Gams, K., Shewale, J., Demian, N., Khalil, K. & Banki, F. (2017, April). *Characteristics, length of stay, and hospital bills associated with severe odontogenic infections in Houston, TX.* [Journal of the American Dental Association, 148(4), 221-229] (https://doi.org/10.1016/j.adaj.2016.11.033)
- Shastri, Y. (2023, April 20). *A Step-by-Step Guide to Federated Learning in Computer Vision.* [V7 Blog] (https://www.v7labs.com/blog/federated-learning-guide#:~:text=Federated%20learning%20(often%20referred%20to,model%20locally%2C%20increasing%20data%20privacy.)
- Song, P. (2023). *Federated Statistical Learning and Distributed Computing* [PowerPoint slides]. BIOSTAT 620, University of Michigan, Ann Arbor, MI.
- Yang, H., Lam, K., Xiao, L., Xiong, Z., Hu, H., Niyato, D., & Poor, H. V. (2022, July 25). *Lead federated neuromorphic learning for wireless edge artificial intelligence.* [Scientific Reports, 12(1), 13540] (https://scite.ai/reports/10.1038/s41467-022-32020-w)

Appendix

Network Diagram

Federated Learning vs. Oracle Summary

Table 3: Federated learning (left) vs. Oracle (right)

Estimate	Std. Error	t value	p value	Estimate	Std. Error	t value	p value
25853.99	1040.99	24.84	0.00	25853.99	1040.99	24.84	0.00
8.83	42.81	0.21	0.84	8.83	42.81	0.21	0.84
-0.15	0.41	-0.37	0.71	-0.15	0.41	-0.37	0.71
-37.72	140.63	-0.27	0.79	-37.72	140.63	-0.27	0.79
-820.87	694.01	-1.18	0.24	-820.87	694.01	-1.18	0.24
293.81	671.01	0.44	0.66	293.81	671.01	0.44	0.66
325.98	672.38	0.48	0.63	325.98	672.38	0.48	0.63
-788.46	695.03	-1.13	0.26	-788.46	695.03	-1.13	0.26
296.05	674.26	0.44	0.66	296.05	674.26	0.44	0.66
372.44	200.33	1.86	0.06	372.44	200.33	1.86	0.06
-816.01	198.38	-4.11	0.00	-816.01	198.38	-4.11	0.00
259.99	283.10	0.92	0.36	259.99	283.10	0.92	0.36
-76.85	282.49	-0.27	0.79	-76.85	282.49	-0.27	0.79
592.70	280.41	2.11	0.03	592.70	280.41	2.11	0.03
-384.31	283.38	-1.36	0.18	-384.31	283.38	-1.36	0.18
-19.94	16.33	-1.22	0.22	-19.94	16.33	-1.22	0.22
28.91	36.56	0.79	0.43	28.91	36.56	0.79	0.43
-25.67	36.03	-0.71	0.48	-25.67	36.03	-0.71	0.48
-17.58	36.12	-0.49	0.63	-17.58	36.12	-0.49	0.63
80.99	37.30	2.17	0.03	80.99	37.30	2.17	0.03
-37.93	36.28	-1.05	0.30	-37.93	36.28	-1.05	0.30