بسم الله الرحمن الرحيم

King Fahd University of Petroleum and Minerals
Deanship of Student Affairs
Vice Deanship for Student Excellence & Success

جامعة الملك فهد للبترول والمعادن
عمادة شؤون الطلاب
وكالة العمادة للتميز والنجاح

Undergraduate Research Office (URO)

Summer Undergraduate Research Experience - SURE203

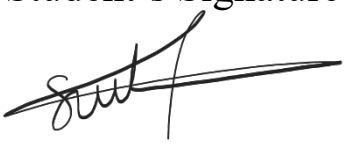
Final Report

Student's Name:	Sultan Abdulsalam Algarbi	**ID:**	201660320
Department:	Information and Computer Science Department	**Level:**	Senior

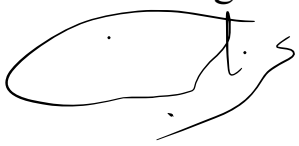
Research Interest	Summer Training
Research Topic	Privacy-Preserving AI Framework using Federated Learning
Advisor	**Dr. Muhamad Felemban** Research Center: Interdisciplinary Research Center for Intelligent Secure Systems

Date: 2021-08-30

Student's Signature



Advisor's Signature



Privacy-Preserving AI Framework using Federated Learning and Linear regression

Sultan Abdulsalam Algarbi
Information and Computer Science Department, KFUPM
Sultanye8@gmail.com

Abstract:

Data plays an important role in machine learning due to its impact on model performance, and there are some practical challenges to sharing data across different parties (clients). One of the challenges is the risk of data leak once the data is shared with third parties. In addition, some laws and regulations prohibit the sharing of some type of data. In this paper, we present a solution that enables data owners to develop a global model without the need to exchange data between parties directly or through a central system. As compared with existing works, this framework is the first work that applies the federated learning concepts with the linear regression problems. The system is built on the parameter server architecture and aims to produce a global model that can be used by several parties to predict the value of a variable based on the values of other variables. We also compare the performance of our framework and the traditional central system, and the comparison results show a great convergence between them.

Keywords:

Federated Learning, Linear regression, data privacy

1. Introduction:

Machine learning, which is a subfield of AI, is widely adopted in many fields, for example, speech recognition, image recognition, traffic prediction, and email spam and malware filtering.

In machine learning (ML), data plays an important role due to its impact on model performance, and in the traditional systems, machine learning models are trained over centralized data that can be collected from different parties (clients), which means the parties share their local data directly (second-party data) or via central endpoint (third-party data). In practice, some types of data cannot be shared for several reasons. The most important of these reasons is the sensitivity of the information to its owner, in addition to the regulations and laws that limit the sharing of some types of data, such as maintaining the privacy and confidentiality of individual information.

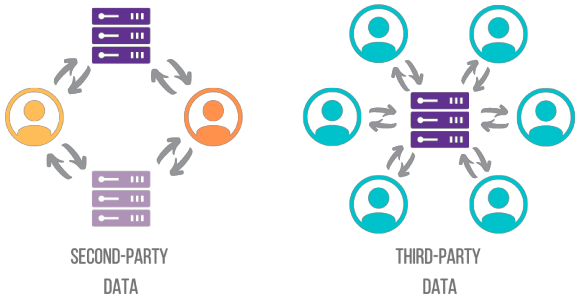


Figure 1: Sharing data in traditional ML frameworks

In the last years, many academic and industry solutions were proposed to effectively address the challenges of data privacy, and one of the most important solutions is the federated learning framework, which was proposed by Google in 2016.

In federated learning, each client trains a sub-model at its site using only local data, and then the clients will share their sub-models in order to reach a consensus on a global model. The global model will be used many times to improve the sub-models. Finally, the performance of the sub-models and the global one will be close to the central model performance in the traditional ML frameworks.

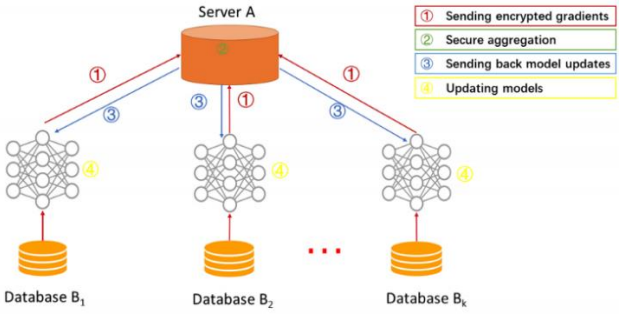


Figure 2: Sharing models in FL framework

The main contributions of this paper can be summarized as follows:

1) We propose a new framework for the global linear regression model that can be built from sub-models trained by several parties.
2) We implement the first linear regression framework based on the federated learning concepts.
3) Our proposed framework can be used by two or more parties.
4) Our proposed framework can be used with any dataset (with some constraints).

The rest of this paper is organized as follows. In section 2, we give a brief overview of related works. We present the details of our solution in Section 3. Experimental results and some analysis are given in Section 4. We conclude the paper in Section 5.

2. Related Work:

This paper is related to federated learning (FL), a new machine learning mechanism that enables to train a joint model on a large corpus of decentralized data owned by different parties while preserving data privacy. In addition, this paper applies the concepts of FL on the linear regression models, which are used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (Y) and a series of other variables (independent variables).

Cheng et. al. [2020] define the meaning of FL as an engineering and algorithm framework to ensure the preservation of data privacy and achieve the confidence of each party dealing with the global artificial intelligence system. Furthermore, they divide the types of FL into three categories: vertical FL, horizontal FL, and federated transfer learning.

Chhikara et. al. [2021] try to discuss a solution that can improve the work environment after Covid-19. they combine the FL with the emotional analysis to create a powerful system that can monitor the emotional behaviors of the users to evaluate their mental health. This system will maintain the privacy of users' data through its use of federated learning, and through this system, the experts will provide the necessary psychological counseling for those who appear mental problems immediately.

Sattler et. al. [2020] state that, in federated learning, the performance of the result of the global modal is low, if the clients' data diverge, and to address this issue, they present clustered FL framework that uses the geometric properties of the federated learning to collect the client population into clusters, which will improve the performance of the result of the global modal.

Ahmed et. al. [2020] present how FL can benefit from using active learning to deal with the unlabeled data and to prove their solution, they applied it with two applications, called natural disaster analysis and waste classification. Hua et. al. [2020] present a new approach that combines federated learning and blockchain, which will provide a mechanism and method for exchanging information necessary to develop a global model while maintaining the privacy of each party's data. Khan et. al. [2020] propose a new framework for Dispersed Federated Learning (DFL) to provide resource optimization, where the distributed learning method provides robustness.

Bonawitz et. al. [2019] developed a scalable federated learning system for mobile devices based on the TensorFlow library which contains a huge number of pre-built software, also describe the overall design of the system, with a clarification of a number of challenges and difficulties facing the researcher in this scope.

Federated learning is a model in which a central model is trained by sharing sub-models of their training results, which are worked out locally within each branch. One of the problems that federated learning faces are that it is relatively slow. So, in order to deal with this problem, Konečný et. al. [2017] present two methods for reducing communication costs between sub-parties and the global central model: Structured updates and the Scheme of updates.

Relying on artificial intelligence and machine learning has become a necessity nowadays when dealing with a huge amount of information and data that requires a great effort to extract accurate statistics and information. One of the areas of life that use these technologies the most is modern health care systems. However, these systems suffer from poor communication and mutual investment in data to develop a more robust health system. Rieke et. al. [2020] discuss the reasons why each health organization does not share its data with others and explore how federated learning can be leveraged to solve this problem.

3. Solution:

3.1 Overview of Multiple Linear Regression:

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables).

Regression helps investment and financial managers to value assets and understand the relationships between variables,

such as commodity prices and the stocks of businesses dealing in those commodities.

The two basic types of regression are simple linear regression and multiple linear regression. Simple linear regression uses one independent variable to explain or predict the outcome of the dependent variable Y, while multiple linear regression uses two or more independent variables to predict the outcome.

The general form of each type of regression is:

Simple linear regression:

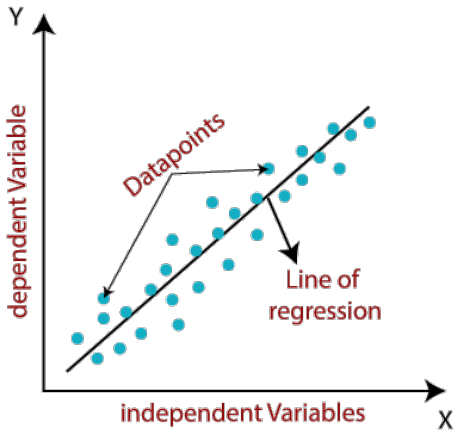
$$Y = a + bX + u$$

Multiple linear regression:

$$Y = a + b_1X_1 + b_2X_2 + b_3X_3 + ... + b_tX_t + u$$

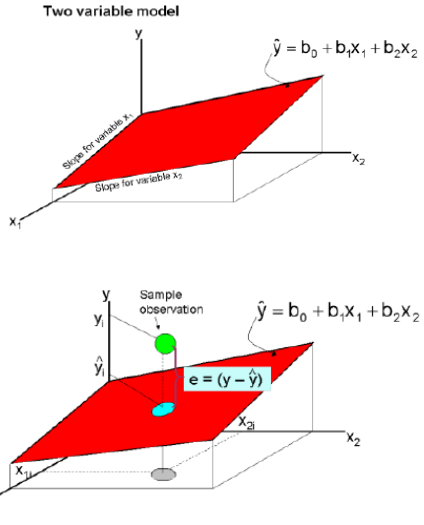
Where:

- Y = the variable that you are trying to predict (dependent variable).
- X = the variable that you are using to predict Y (independent variable).
- a = the intercept.
- b = the slope.
- u = the regression residual.



Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a straight line (linear regression) that best approximates all the individual data points. In multiple regression, the separate variables are differentiated by using subscripts.

Multiple Regression Model (Two variable model)



3.2 Protocol of Model Training:

To train the linear regression model with federated learning concepts, we need to protect the row data (Xs, Y) for each party (client). These data are stored locally by each party and are forbidden to be transmitted.

The main steps of our proposed framework:

1) Training the local dataset of each client on his site.
2) Sharing the linear regression models of all the parties with the central server.
3) Building the global model using the sub-models.
4) Sending back the global model to the parties.
5) Updating the local model of each party.

Steps 2 to 5 will be usually iterated a lot of times until a maximum iteration number is reached or some convergence conditions are satisfied.

3.3 System Architecture:

We aim to build a federated learning framework that can be used to handing the linear regression issues without violating data privacy. Figure 3 shows the system architecture of our proposed solution.

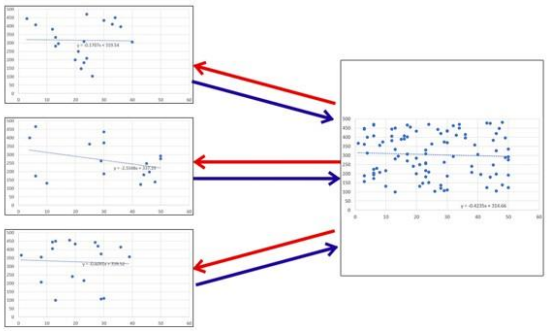


Figure 3: system architecture

Each party uses the least-square method to estimate the regression coefficients of the multiple regression model.

Figure 4 shows the way of representing the dependent variables and the independent variable for each row in the dataset.

y	x_1	x_2	. . .	x_k
y_1	x_{11}	x_{12}	. . .	x_{1k}
y_2	x_{21}	x_{22}	. . .	x_{2k}
.	.	.	. . .	.
.	.	.		.
y_n	x_{n1}	x_{n2}	. . .	x_{nk}

Figure 4: Xs and Ys representation

The least-square function is given by:

$$L = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j x_{ij} \right)^2$$

The least-square estimate must satisfy the following:

$$\left. \frac{\partial L}{\partial \beta_0} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0$$

$$\left. \frac{\partial L}{\partial \beta_j} \right|_{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k} = -2 \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{k} \hat{\beta}_j x_{ij} \right) x_{ij} = 0, \quad j = 1, 2, \ldots, k$$

The least-square normal equations are the least-square estimators of the regression coefficients:

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik} = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \hat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \hat{\beta}_2 \sum_{i=1}^{n} x_{i1} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{i1} x_{ik} = \sum_{i=1}^{n} x_{i1} y_i$$

$$\cdots \quad \cdots \quad \cdots$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_{ik} + \hat{\beta}_1 \sum_{i=1}^{n} x_{ik} x_{i1} + \hat{\beta}_2 \sum_{i=1}^{n} x_{ik} x_{i2} + \cdots + \hat{\beta}_k \sum_{i=1}^{n} x_{ik}^2 = \sum_{i=1}^{n} x_{ik} y_i$$

The solution to the normal equations is the least-square estimators of the regression coefficients.

4. Experiments:

4.1 Settings:

Dataset: To set up this framework, we have to select any dataset with the following constraints:

1. The dataset file extension should be "*.csv*".
2. The first row in the file represents the labels of the features.
3. The remaining rows represent the samples.
4. The type of each element in the dataset should be a number (integer, float, double) with no empty element.
5. From the first column to the column before the last, the data represent the independent variables [Xs], and the last column represents the dependent variable [y].

Federated Clients (parties): this framework has the ability to be used by any number of clients (parties), and to set up this experiment, the user should enter the number of clients (parties), wherein this experiment, the user enters the dataset as one block, then the framework will distribute the dataset to the clients. In the real-world implementation of Federated learning, each client has its dataset in isolation. The shard creation step in the framework only happens in the experiments.

Training and Testing samples: the user has the ability to enter the percentage of the testing samples, then the framework will split the dataset into two sub-datasets, one for the training, and the other for the testing. The training samples will be sharded to sub-datasets, where each sub-dataset will be associated with one client. The testing samples will be used to test the federated model and the centralized model.

Training client datasets: each client trains a sub-model at its site using only local data, and then shares its sub-model with the central server to make its contribution to the global model. The global model will be used many times to improve the sub-models. The sub-model of each client contains the values of the estimated slope coefficients and the y-intercept.

Our Experiment setup:

- **dataset name**: wire_pull_strength
- **number of clients** = 4
- **testing size** = 20%

4.2 Results:

In this section, we present the experimental results. Table 1 shows the percentage error of the global model (Using federated learning) and the central model (Using traditional central server).

Framework	Model	% Error
Federated Learning	Global	4.41 %
Traditional (without FL)	Central	4.62 %

Table 1: The percentage error results

Figure 5 shows the matrix of scatter plot of the experiment:

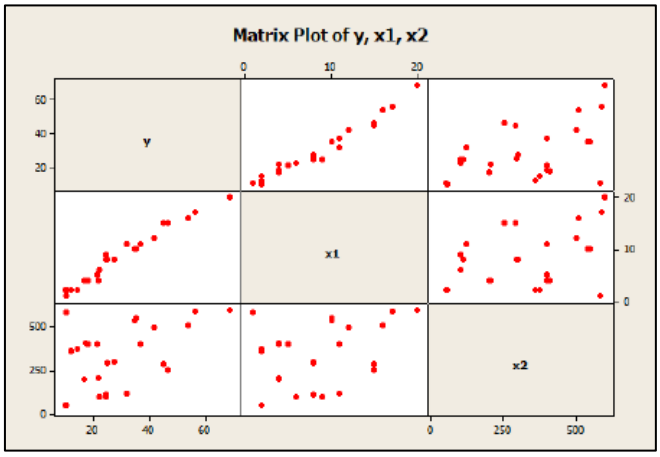


Figure 5: matrix of scatter plot

5. Conclusions:

Federated Learning is becoming an important topic for the researcher and industrial communities due to its ability to preserve data privacy. In the coming years, many systems in machine learning will be built based on federated learning. In this paper, we propose a new framework for the global linear regression model that can be built from sub-models trained by several parties, and we implement the first linear regression framework based on the federated learning concepts. Moreover, our proposed framework has the ability to be used by two or more parties with any dataset (with some constraints).

References:

[1] Yong Cheng, Yang Liu, Tianjian Chen, and Qiang Yang. 2020. Federated learning for privacy-preserving AI. Commun. ACM 63, 12 (December 2020), 33–36. DOI:https://doi.org/10.1145/3387107

[2] P. Chhikara, P. Singh, R. Tekchandani, N. Kumar and M. Guizani, "Federated Learning Meets Human Emotions: A Decentralized Framework for Human–Computer Interaction for IoT Applications," in IEEE Internet of Things Journal, vol. 8, no. 8, pp. 6949-6962, 15 April15, 2021, doi: 10.1109/JIOT.2020.3037207.

[3] F. Sattler, K. -R. Müller and W. Samek, "Clustered Federated Learning: Model-Agnostic Distributed Multitask Optimization Under Privacy Constraints," in IEEE Transactions on Neural Networks and Learning Systems, doi: 10.1109/TNNLS.2020.3015958. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9174890&isnumber=6104215

[4] L. Ahmed, K. Ahmad, N. Said, B. Qolomany, J. Qadir and A. Al-Fuqaha, "Active Learning Based Federated Learning for Waste and Natural Disaster Image Classification," in IEEE Access, vol. 8, pp. 208518-208531, 2020, doi: 10.1109/ACCESS.2020.3038676. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9261337&isnumber=8948470

[5] G. Hua, L. Zhu, J. Wu, C. Shen, L. Zhou and Q. Lin, "Blockchain-Based Federated Learning for Intelligent Control in Heavy Haul Railway," in IEEE Access, vol. 8, pp. 176830-176839, 2020, doi: 10.1109/ACCESS.2020.3021253 URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9184854&isnumber=8948470

[6] L. U. Khan, M. Alsenwi, I. Yaqoob, M. Imran, Z. Han and C. S. Hong, "Resource Optimized Federated Learning-Enabled Cognitive Internet of Things for Smart Industries," in IEEE Access, vol. 8, pp. 168854-168864, 2020, doi: 10.1109/ACCESS.2020.3023940. URL: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=9195820&isnumber=8948470

[7] Keith Bonawitz, Hubert Eichner, Wolfgang Grieskamp, Dzmitry Huba, Alex Ingerman, Vladimir Ivanov, Chloe Kiddon, Jakub Konečný, Stefano Mazzocchi, H. Brendan McMahan, Timon Van Overveldt, David Petrou, Daniel Ramage, Jason Roselander

[8] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, Dave Bacon

[9] Rieke, N., Hancox, J., Li, W. et al. The future of digital health with federated learning. npj Digit. Med. 3, 119 (2020). https://doi.org/10.1038/s41746-020-00323-1