ML - Federated Learning - Application in life Insurance industry

Abstract - In the healthcare industry where medical insurance providers are competing with each other to acquire more and more customers, evaluating customers' application to assign a risk level is of prime importance. This helps in formulating the policies and the premium that a customer needs to pay. In order to work on this the insurance companies must share their data which is highly susceptible of being stolen and misused against them by their corporate rivals. Federated learning (FL) is introduced which works in a distributed fashion without sharing and accessing the actual data. In this paper, we will try to analyze and exploit this concept to provide risk assumptions in the Healthcare and Life Insurance Industry.

Important Keywords - Federated learning, Risk level prediction, Keras sequential model, Quadratic Weighted Kappa Store.

1- Introduction - Many organizations have started employing machine learning for risk predictions. These predictors may get biased if the company serves majorly only a particular type of customer. So the model will fail when predicting risk level for a different type of customer. One possible solution is to collect data from different companies and train the model. But it is highly unlikely that one company will exchange sensitive customer data with other companies. This is where federated learning comes in, contributing to make the risk predictors more robust, as the models can be trained across servers of different organizations while protecting data privacy.

The concept of FL is in a way a more decentralized and privacy protecting method of Machine learning. The basic problem it solves is allowing created models to earn knowledge from the data with actually accessing or exposing any section of it. Organizations train Models in house and then share the trained models to a centralized location. The centralized model then combines the entities received and shares it with the partners for the next iterations of local training. This type of FL is called cross-silo (where silo stands for organization). The other type of FL is called cross-device. These devices can be computationally low-end such as smartphones. However, for some high data driven tasks FL avoids training on these low-capacity devices.

FL can be further classified into two types based on the sample space and features namely horizontal FL and vertical FL. In horizontal FL (HFL) Fig 1, different clients participating in the learning process have overlapping or the same features but entirely different sample spaces. On the other hand, in vertical FL, different clients have different features but the same sample space.

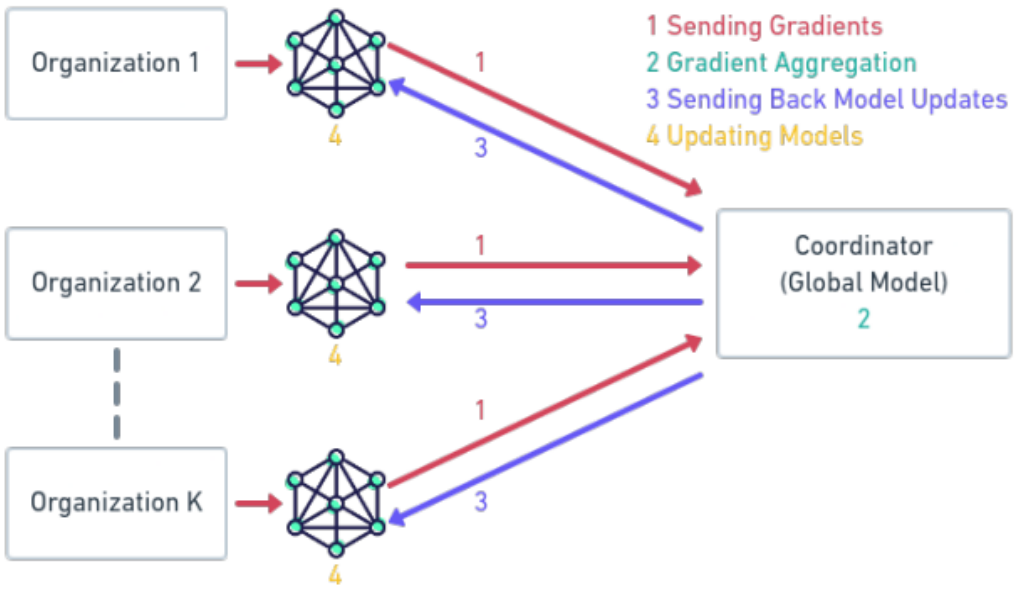

Fig 1 . Architecture of Horizontal FL.

The problem we've picked is a type of Horizontal FL. We'll go with literature in section 2. The methodology of this white paper is discussed in section 3. The conclusion in section 4.

2- Literature Review - The life insurance companies do a lot of underwriting tasks to assess their customers and issue the policies. This underwriting task is what helps insurance companies to decide the premiums and policies which they can offer to a particular customer.

This is accomplished through several medical examinations whose results need to be submitted to the insurance company. Through this data companies analyze the risk profile of the customer and based on this information, companies can decide whether to accept or reject the application and also calculate the premium of the policy. This underwriting task is essential for the companies to maintain a repo and an advantageous position in the competitive market.

Numerous machine learning methods have been developed to replace complex manual calculations of the risk profiles. This improves the underwriting process and increases the pace of it. Algos like Multiple Linear regression, Artificial Neural Network, REPTree, and Random Tree have been tested for risk prediction and have been significantly fair in results.

Let us propose the use of Federated Learning across the organizations to tackle the problem of collaboration and make a more robust model thus increasing the scope of growth for companies while improving the underwriting process to maintain customer satisfaction and acquire new customers on their journey.

3- Methodology Preview - Let us exploit Horizontal FL with a similar set of features (age, height, weight, insurance history, medical history etc) in the local data sets and different sample spaces for each organization. Horizontal main advantage is its ability to preserve privacy for each organization. We have a dataset which will be used for analysis. In HFL, the dataset is divided among different clients and used to model training with a varying number of clients and distribution. We evaluate the proposed method using Quadratic Kappa Store. The proposed method is also compared to a centralized approach, where the complete data is offloaded to a central server that executes the model training. Fig.2 shows the entire methodology in the form of a flow chart.

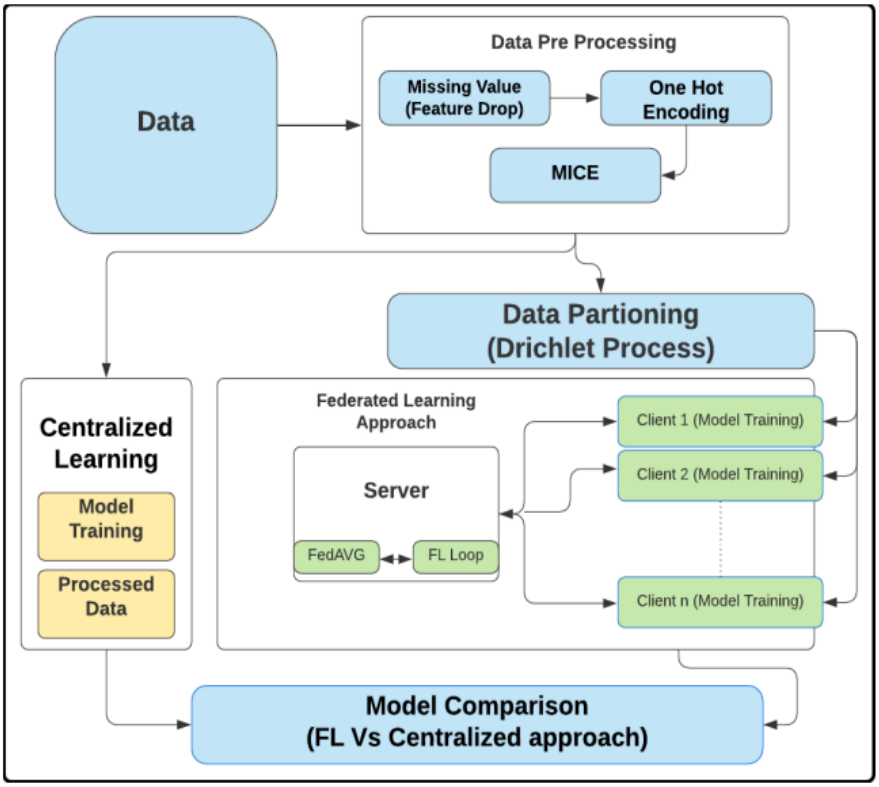

Fig.2: Flowchart of the proposed methodology.

Let us explain each and every block for more clear understanding.

1) *Missing Value (Feature Drop)* - This block itself explains the relevance during modeling. All features which are being used inside HFL, if they have more than 40% of the missing data then such features are discarded for consideration, as filling these many values make data more random and less real.
2) *One Hot Encoding* - This block represents the categorical feature conversion. There are different kinds of variables that cannot be directly fed into the ML and hence require hot encoding which they are converted into a form that can be easily understood by ML. Eg. Highschool, UG, Ph.D. are converted into 1,2,3 and 4 while using these variables during HFL. Thus numerical conversion is what we call hot encoding.

3) *MICE* - There may be a situation where during modeling the dataset might have some missing values which must be added for the effective training of the model. To fill such missing values the technique is what we call MICE (Multivariate Imputation by Chained Equations). MICE creates a chain where predictions of missing data are affected by other missing data values. MICE trains data on non-missing values using a regression model to predict missing data.
4) *Data Partitioning* - In a real-world scenario, the distribution of data across the clients is expected to differ significantly. Thus, we have to partition the data in the dataset among different clients in a manner that preserves the statistical properties. As shown in Fig 2, the Dirichlet Process is used, where the concentration parameter (α) is taken into consideration. The α is mainly responsible for how different classes of risk levels will be distributed to each client so that a ratio of class distribution is maintained among these clients.
   For $\alpha \rightarrow 0$, the distribution is obtained very sparsely, with the distribution of risk levels varying highly between clients. As α increases, the distribution starts looking similar on clients, and $\alpha \rightarrow \infty$ results in almost the same distributions among the clients and these distributions are also identical to the base distribution. Different values of α lead to different data distribution and for the proposed work following values of α are taken into consideration.

Fig.3 - Distribution of classes for various values of α.

5) *Federated Learning approach* - In fig 2 the federated learning approach block represents the proposed machine learning technique. The method adopts a distributed approach to train the model. There is one server and multiple distributed clients. Each client contains its private data which is used for training their local model. The orchestration between the client's model and server model is done through the FedAvg algorithm. The shared global model is hosted on the central server. It is during this phase the Quadratic

Weighted Kappa score is calculated. Just for sample purpose let us have a look at equation: where

$$\kappa = \frac{\sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}P_{ij} - \sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}P_iP_j}{1 - \sum_{i=1}^{I}\sum_{j=1}^{I} w_{ij}P_iP_j}$$

$$w_{ij} = 1 - \frac{(i-j)^2}{(N-1)^2}$$

4- Conclusion - Federated Learning allows multiple organizations to come together to create a predictive model without actually sharing the data. In the case of life insurance risk prediction, it presents an opportunity for multiple insurance companies to come together and create a model which is robust and more capable.

To approach HFL start with a dataset. The dataset had many missing values, so as the first step, features having more than 40% missing values were removed. Then one hot encoding was used to convert non-numerical categorical features to numeric features. Next, the remaining missing values were filled using Bayesian Regression MICE. In a FL approach, every client consists of its own data and trains its model locally. The statistics of the dataset likely vary between clients. To simulate this, the Dirichlet Process is used to partition the dataset among the clients to cover a spectrum of similarity.

Fig.5 - Sample FedAvg Learning Curves.

The obtained analysis shows it is possible to perform the task of life insurance risk prediction using federated learning, and the performance is not that far off from that of the traditional machine learning approaches. However there are few limitations of the research we conducted. Data is divided equally among clients and therefore we directly average the local model updates to get global model updates. But in a practical scenario we can have clients holding unequal data, therefore during updating the global model update we can use the weighted average in proportion to the amount of data each client holds.