國立政治大學風險管理與保險學系
碩士論文

聯邦學習：肺癌存活率預測

指導教授：謝明華教授

研究生：劉源 撰

中華民國 111 年 1 月

DOI:10.6814/NCCU202200065

摘要

在數據保護愈發嚴格的情勢下，保險公司在遵守數據安全保護的前提下，如何利用更多的數據對於癌症險的出險、費率進行進一步預測。本文探討了一種解決企業之間數據不能相互傳輸的方式：聯邦學習。本文透過預測肺癌的存活率，比較了聯邦學習和傳統機器學習的評估效果。結果發現，聯邦學習在數據不能出本地的情況下，依舊可以達到和傳統機器學習類似的效果。因此，本文認為，聯邦學習可以在保險公司的費率、出險率的預測上提供一種新的思路，幫助保險公司克服所面臨的數據量不足，受到法規限制等問題。

關鍵詞：聯邦學習、肺癌、數據孤島

DOI:10.6814/NCCU202200065

Abstract

Under the situation of increasingly strict data protection, it’ s important for insurance companies to further predict the risk and rate of cancer insurance with more data. This paper discusses a way to solve the problem that data cannot be transmitted between enterprises—Federated learning. By predicting the survival rate of lung cancer, this paper compares the effects of federal learning and traditional machine learning. The results show that federated learning can achieve the same effect as traditional machine learning when the data must stay in local. Therefore, this paper shows that under the restriction of laws and regulations federal learning can provide a new direction in the prediction of survival rate for insurance companies to overcome the problems of insufficient data.

Keywords：Federal learning、lung cancer、 data island

DOI:10.6814/NCCU202200065

目錄

DOI:10.6814/NCCU202200065

圖次

DOI:10.6814/NCCU202200065

表次

DOI:10.6814/NCCU202200065

第一章　緒論

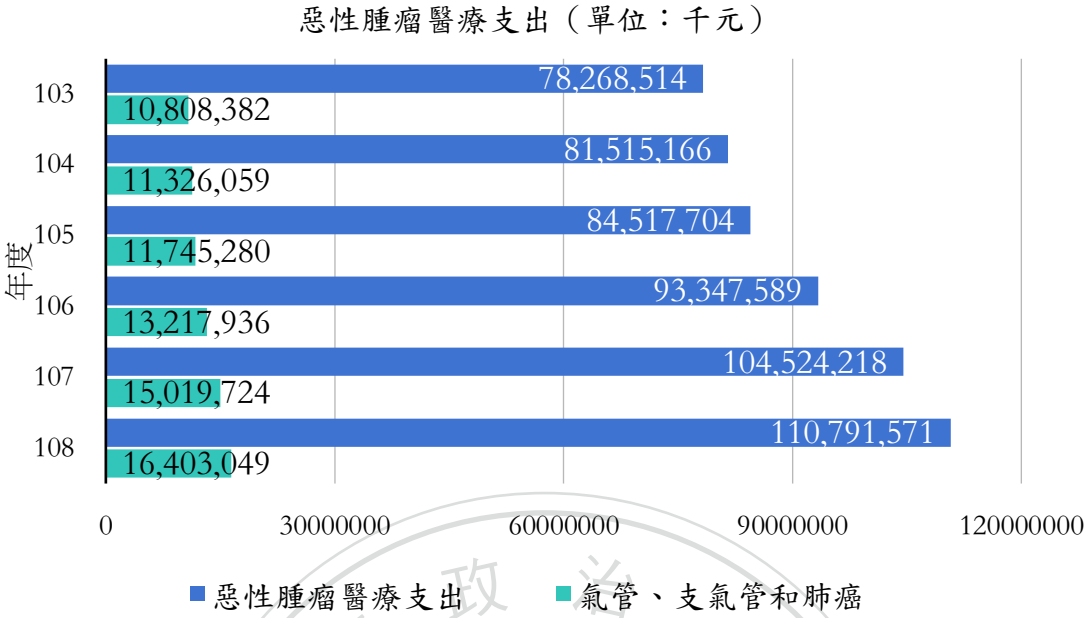
第一節　研究動機

2020 年，國際癌症研究機構(International Agency for Research on Cancer,簡稱 IARC)發布的最新關於癌症的報告中，全球新發癌症 1929 萬例，死亡病例高達 996 萬例。雖然乳腺癌在 2020 年首次超過了肺癌的發病率，但是肺癌仍然是全球死亡率最高的癌症，死亡人數 180 萬，近乎第二名結直腸癌 94 萬的兩倍。全球範圍內，城市工業化和經濟發展帶來了大氣污染和室內環境污染都增加了肺癌的發病機率。人類發展指數(Human Development Index，HDI):一種用於衡量各國人類發展的綜合指標，包括了健康長壽、受教育水平和生活水平，國際癌症研究機構發現，HDI 指標高的地區肺癌的發病率與死亡率都較高。隨著現代化發展，肺癌的發病率逐漸提升，對於保險公司來說，如何預測這種疾病的存活率，存活年限是至關重要的。

在台灣，根據衛生福利部中央健康保健署的資料，每年用於惡性腫瘤的醫療支出也在逐年增加，肺癌支出佔惡性腫瘤醫療支出的 15%，醫療費用高居榜首，如圖 1-1 所示。同時，根部衛生部健保署 108 年度惡性腫瘤醫療支出資料，截止 2019 年，罹患癌症的病人每人平均醫療費用高達 228,013 元，平均藥費達到 105,804，醫療費用 5 年年平均增長達到 9.7%。

DOI:10.6814/NCCU202200065

圖 1　　惡性腫瘤支出費用



資料來源：衛生福利部中央健康保健署

根據保發中心的統計資料顯示，2019 年度壽險業死亡給付金額為新台 117,230,652 千元。就死亡原因分析，惡性腫瘤的理賠給付高居首位，人數佔死亡理賠的 41.71%，如表 1-1 所示。惡性腫瘤 10 年來新確診人數持續上升，男性增加 3 成，女性則翻倍，除此之外，肺癌存在年輕化趨勢，男女性癌症發生中位數下降均超過 3 歲。因此，越早的癌症險規劃和定期檢查，可以幫助被保險人在面對突發風險時，可以有更多的保障，且保費相對便宜。

表 1　　壽險被保險人死亡率原因

	人數	佔比%	給付金額（單位：千元）	佔比%
惡性腫瘤	84,690	41.73	48,174,138	41.09

DOI:10.6814/NCCU202200065

	人數	佔比%	給付金額（單位：千元）	佔比%
心臟病	25,823	12.72	14,558,927	12.42
意外之災害	16,708	8.23	12,410,308	10.59
肺炎	14,650	7.22	7,137,200	6.09
腦血管疾病	8,540	4.21	4,828,535	4.12
肝硬化	3,447	1.70	1,757,828	1.50
腎炎	1,892	0.93	864,538	0.74
呼吸器官結核病	780	0.38	404,008	0.34
支氣管炎	545	0.27	216,018	0.18
其他	45,863	22.60	26,879,152	22.93

資料來源：財團法人保險事業發展中心，108 人壽保險業務統計年報

癌症的發病率逐漸增加，治療費用昂貴，引起了民眾對於癌症高風險的意識逐漸提高。1979 年，台灣發佈了第一張終身癌症險，並且由較為單一的理賠方式逐漸發展出更多樣的商品，包括帳戶型、倍數型、還本型和一次給付型等等商品，產、壽險公司紛紛推出防癌險。

隨著治療技術的進步，癌症的治療方式更佳多元化，但是醫療費用也越發昂貴，存活率上升。但是對於保險業者來說，癌症險理賠額度較高，一旦保費收取不足，損失率攀升，很有可能造成保險公司大量虧損，保險公司只能降低保費，或者調整理賠範圍以因應目前的環境。

癌症的治療費用高，對於家庭造成了非常沈重的經濟和心理負擔。作為轉嫁風險的保險公司，存活率和醫療費用的上升也造成了經營的壓力。因此，癌症的存活率對於保險公司來說是非常重要的保費計算依據。而肺癌作為最高死

DOI:10.6814/NCCU202200065

亡率的癌症，且治療費用最為昂貴的疾病，肺癌的存活率則是也是對於保險公司非常重要的指標，因此本文希望通過機器學習的方法，進行肺癌生存率的預測，給予保險公司精算上的一些幫助。

傳統的機器學習方法往往需要將各家醫院的資料彙整集中，在進行建模。但是，由於個人資料隱私的保護，法規限制等等，個人資料的使用往往是受到限制。而對於癌症來說，目前癌症醫療庫的建構還不夠完善，各家醫院如果只採用自己的資料建立模型，會由於資料量過少，造成預測不準確的現象。而本文通過聯邦學習解決該問題。聯邦學習是由 Google 率先提出的一種分布式學習，可以避免資料的洩露，通過本地樣本去訓練模型，而無需分享數據，數據不會離開用戶本地，從而可以解決數據隱私、安全性等等問題。因此，除了傳統的機器學習方法之外，本文希望模擬不同醫院資料不能互相傳輸的環境，通過聯邦學習的方式進行訓練，並且比較聯邦學習和傳統機器學習的模型評估效果是否存在差異。

第二節　研究架構

本研究主要通過傳統機器學習和聯邦學習預測肺癌存活率，並比較兩種方式之間是否存在差異，全文共有四個章節，概述如下：

一、緒論

簡述透過聯邦學習預測肺癌死亡率的背景和動機以及架構。

二、機器學習方法介紹

首先介紹了後文中會使用的傳統機器學習的方法，包括決策樹，隨機森林，knn 等方法，第二節會介紹聯邦學習的概念，並且介紹本研究會使用的聯邦學習模型：SecureBoost。

三、實證研究

該章節通過 SEER 資料庫的 2010-2016 年的肺癌資料，第一節論述了本研究如何進行資料前處理，刪除缺失值並對類別變數進行處理，通過傳統機器學習方法預測 2 年和 5 年肺癌存活率；之後模擬不同醫院不能互相傳輸資訊的場景，將樣本分群，進行聯邦學習預測。最後整理兩者結果，比較傳統機器學習和聯邦學習的效果差異。

DOI:10.6814/NCCU202200065

四、結論

針對本文研究結果進行探討，研究發現：聯邦學習下學習效果和傳統積極學習效果相差不大，即聯邦學習可以在保護本地數據的同時進行訓練。同時並檢討本次研究的不足之處，給予後續研究方向的建議。

DOI:10.6814/NCCU202200065

第二章　機器學習方法介紹

在大數據時代，數據資料的存儲量意味著人們可以有更多的資訊，但是實際上，大量的人工無法辨識、識別或者複雜的數據，光靠人工的力量是無法得到數據的想提供給我們的資訊，機器學習就是通過算法代碼的方式，從大量的數據中獲得人們想要知道的資訊。同時，由於環境在變化，人們每天都會收集到新的數據，反應的資訊可能隨時在變換，機器學習可以很快地適應不斷變更的龐雜的數據，幫助處理人工難以處理的問題。機器學習包括了分類和回歸問題，本文主要是探討了關於肺癌存活率的分類問題。

第一節　傳統機器學習介紹

機器學習是透過不同的演算方法將數據進行訓練，適用於回歸或者分類問題，可以用於對新的資料進行預測。本研究主要要研究的是癌症存活率的預測，並且訓練的數據集擁有標籤，屬於無監督式的分類問題，下文介紹 6 種本研究會用到的機器學習分類模型。

一、　邏輯式回歸（Logistic Regression）

邏輯式回歸往往用來估計某一樣本點屬於某一類別的可能性，採用了對數函數 sigmoid function。

Sigmoid 方程如下所示：

$$\sigma(t) = \frac{1}{1 + \exp(-t)}$$

DOI:10.6814/NCCU202200065

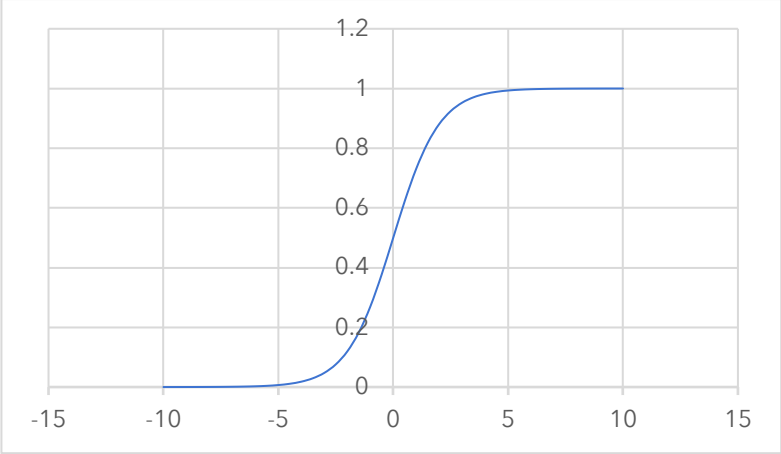


圖 2 Sigmoid function

而邏輯式回歸則透過線性回歸找出參數θ，並預測出 y 值，將 y 值帶入 sigmoid function，得到 0 到 1 之間的概率。

$$\hat{p} = h_\theta(x) = \sigma(\theta^T \cdot \boldsymbol{x})$$

新樣本點的預測值則可以通過上述訓練好的模型進行預測，如果估計的可能性超過 0.5，則預測為類別 1，反之為 0。

$$\hat{y} = \begin{cases} 0 \ if \ \hat{p} < 0.5 \\ 1 \ if \ \hat{p} \geq 0.5 \end{cases}$$

二、 決策樹(Decision tree)

決策樹是一種可以同時用於分類和回歸問題的演算法，可以處理非常複雜的數據資料。是透過選取特定特徵，並以該特徵來切割樣本。而如何劃選取特徵和不同的劃分域值，則是透過 gini 不純度或者熵進行判定：

Gini 不純度(Gini impurity)的公式如下：

$$G_i = 1 - \sum_{k=1}^{n} p_{i,k}^2$$

熵(Entropy)的公式如下

$$H_i = -\sum_{\substack{k=1 \\ p_{i,k} \neq 0}}^{n} p_{i,k} \log_2(p_{i,k})$$

$p_{i,k}^2$為決策樹第 i 個節點，分類為類別 k 的樣本佔該節點總樣本數量的比例。

DOI:10.6814/NCCU202200065

三、 隨機森林(Random Forest)

隨機森林是由很多不同的決策樹所組成的學習器，這種方式又稱為集成學習（Ensemble Method），通常會透過 bagging 的方式進行。集成學習是集合多個模型並匯總他們的預測結果，通常來說，多個模型的預測結果往往會比單一的模型預測的要更準確。

隨機森林就是透過多棵決策樹模型抽取不同的樣本進行預測，得到最終的結果。通常來說抽樣的方式分為 Bagging（Bootstrap Aggregating）和 Pasting。Bagging 採取的是抽取後放回的方式，也就是說樣本是可能重複被抽取的。Pasting 則相反，樣本是不可以重複抽取的。

每一棵決策樹做出預測之後，則採用投票（Voting）的方式決定最終的結果。投票也分為 Hard voting 和 Soft voting 兩種，hard voting 是常見的少數服從多數的投票方式；而 soft voting 則是通過模型輸出的分類概率的平均值進行預測。

四、 XGBoost

Boosting 也是一種集成學習的方式，通過提高預測錯誤的資料的權重，優化對錯誤分類資料的預測，提升模型的訓練結果。而 XGBoost 則是保留上一棵決策樹的預測結果，加入新的模型進行訓練，修正上一棵樹的錯誤預測，從而提升預測結果。

五、 支持向量機（Support Vector Macine）

支持向量機是為了找到一個決策邊界(decision boundary) 可以使得兩種不同類別的資料之間的邊界(margins) 最大化。可以分為 Soft Margin Classification 和 Hard Margin Classification。 Hard Margin Classification 要求分類邊界內不能存在樣本點，也就是資料量必須線型可分。而 Soft Margin Classification 則是允許樣本點出現在邊界之內，在誤差和分類準確率之間尋找平衡。

對於樣本 x 的分類問題，尋找到一個決策方程$w^T x + b$，若$w^T x + b$<0，則分類為 0，否則為 1：

DOI:10.6814/NCCU202200065

$$\hat{y} = \begin{cases} 0 \ if \ \boldsymbol{w}^T\mathbf{x} + \boldsymbol{b} < 0 \\ 1 \ if \ \boldsymbol{w}^T\mathbf{x} + \boldsymbol{b} \geq 0 \end{cases}$$

同時找到兩條邊界$\boldsymbol{w}^T x + b = 1$和$\boldsymbol{w}^T x + b = -1$，如果要使得邊界的距離最大，也就是最小化$\|w\|$。定義 t：

$$t = \begin{cases} -1 \ if & y = 0 \\ 1 \ if & y = 1 \end{cases}$$

本文簡述 Hard Margin Classification 的最佳化方程：

$\underset{w,b}{\text{minimize}} \ \frac{1}{2}\boldsymbol{w}^T \cdot \boldsymbol{w}$ 且服從 $t(\boldsymbol{w}^T x + b) \geq 1$

六、 kNN（k-Nearest Neighbors ）

kNN 是一種廣泛應用的機器學習方式，概念也非常的簡單。對於新樣本點的預測結果解釋透過距離該樣本點最接近的 k 個鄰居的標籤進行投票或者計算平均數得到該點的預測結果。

常用的距離計算公式有，歐幾里德距離：

$$Distance = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$

曼哈頓距離：

$$Distance = \sum_{i=1}^{n}|x_i - y_i|$$

明可夫斯基距離：

$$Distance = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{\frac{1}{p}}$$

在分類問題中，如何決定該模型的優劣，也有很多判定的標準。首先先介紹混淆矩陣的概念(Confusion Matrix)。假設偵測信用違約的狀況下，實際情況下分類有 Positive（有違約）和 Negative（沒有違約）兩類，在模型預測後就會有混淆矩陣所示的四種結果：實

DOI:10.6814/NCCU202200065

際為 Negative 並且預測正確（Negative），實際為 Negative 並且預測錯誤（Positive），實際為 Positive 並且預測正確（Positive），實際為 Positive 並且預測錯誤（Negative）。

表 2　混淆矩陣

		預測結果	
		Negative	Positive
實際情況	Negative	TN True Negative	FP False Positive
	Positive	FN False Negative	TP True Positive

下文介紹五種判定模型優劣的標準：

1. Accuracy 為整個模型的準確率：

$$\text{Accuracy} = \frac{TN + TP}{TN + FP + FN + TP}$$

2. Precision 指預測為違約的人中，實際違約的比例：

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. Recall 也可以成為是 True Positive rate (TPR)，即判定為實際違約的人中判定為違約的比率：

$$\text{Recall} = \frac{TP}{FN + TP}$$

4. F1

$$\text{F1} = \frac{2}{\frac{1}{precision} + \frac{1}{recall}}$$

5. ROC 曲線及 AUC(Area under curve)

ROC 曲線橫軸為 FPR(False positive rate),即模型預測違約的人中實際沒有違約的人數，而縱軸為 recall,對角線為基準線，表明模型沒預測力。當 ROC 曲線越靠近座標軸左上角，說明模型的預測效果越好。AUC 為 ROC 曲線下方面積，取值 0-1 之間，越接近 1，模型預測結果越好。

DOI:10.6814/NCCU202200065

第二節　聯邦式學習介紹

大量的數據和電腦運算能力支撐著了機器學習的發展。雲端技術和大數據時代支撐了機器學習發展的兩個要素。但是隨者各個國家對於數據隱私保護的增加，例如歐盟在 2015 年 5 月 25 日 發布了《一般資料保護規範》（General Data Protection Regulation），如果資料所有者不同意，那麼禁止直接參與自動做出決策的模型。這些規定增加了企業運用數據的壓力，再進行資料數據運用的時候隨時有可能觸犯到這些法規。除此之外，用戶的數據保護意識上升，不願意上傳自己的數據，或者企業與企業之間的數據可能會存在壁壘，譬如醫院想要做關於疾病發生率的研究，但是其實醫院之間的數據是不能離開本地進行傳輸，不然會觸犯法規。為了解決數據壁壘問題，聯邦式學習因應而生。

聯邦式學習是一種分布式學習，通過不同的伺服器或者邊緣設備的本地樣本去訓練模型，而無需分享數據，數據不會離開用戶本地，從而可以解決數據隱私、安全性等等問題。

一、聯邦學習的定義

在 Yang, Liu et al. (2019)中假設有 N 個客戶端，他們的數據為 $\{ D_1, D_2, \ldots, D_N \}$，傳統機器學習就事將這 N 的數據集合起來為 $D_{sum}=D_1 \cup D_2 \cdots \cup D_N$，並通過集中後的的數據訓練模型 M_{sum}。聯邦學習，則是讓每個客戶端訓練自己的數據 D_i，建立模型 M_{fed}。並且我們得到這兩種模型的性能量度（如正確率、Recall 等）分別表示為 V_{sum} 和 V_{fed}。假設存在任意非負數 δ，狹義的聯邦學習定義的正確率可以滿足下式：

$$|V_{sum}-V_{fed}|<\delta$$

在狹義的聯邦學習定義下，相比起傳統機器學習，其訓練產生的效果和傳統機器學習相近。

而在廣義的聯邦學習定義下，聯邦學習具有 δ 的性能量度表示為：

$$V_{sum}-V_{fed}<\delta$$

DOI:10.6814/NCCU202200065

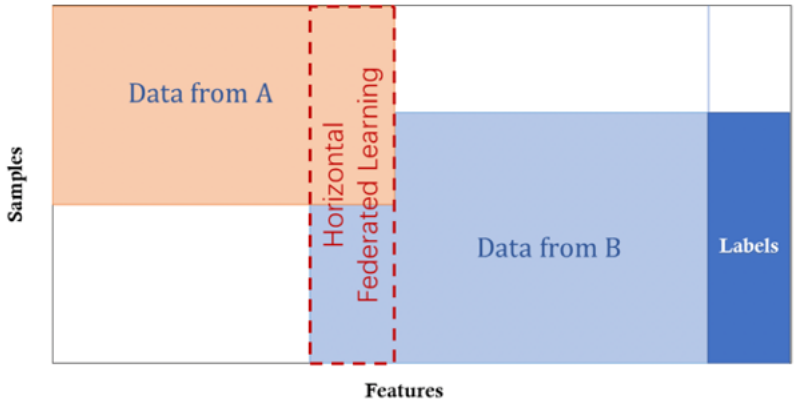
在廣義上，V_{fed} 的取值可以是大於 V_{sum}- δ 的任意值。這主要是如果有多方參與方，他們的數據不平衡，部分參與方的數據質量低，傳統機器學習的效果可能不佳。但是聯邦學習在開始訓練之前，會透過檢測，將異常的客戶端資料剔除，從而產生更好的學習效果；

同時，作者還將聯邦式學習分為分下列三類：橫向聯邦學習、縱向聯邦學習和聯邦遷移學習。

二、聯邦學習的分類

1. 橫向聯邦學習

橫向聯邦學習是基於相同的特徵，可以表示為：$X_i=X_j$，$Y_i=Y_j$，$I_i \neq I_j$，對於任意不同兩家企業 i 和 j，採用相同的特徵，但是企業所提供的樣本不同。橫向聯邦學習按照樣本進行劃分，使得模型的數據量增加。橫向的聯邦學習通常用於跨設備端的場景，目前已經實現的模型，包括了線型模型、GBDT 模型、卷積神經網路模型等等。實際上，使用梯度下降的迭代優化模型基本都可以採用橫向的聯邦學習。



(a) Horizontal Federated Learning

圖 3　橫向聯邦式學習

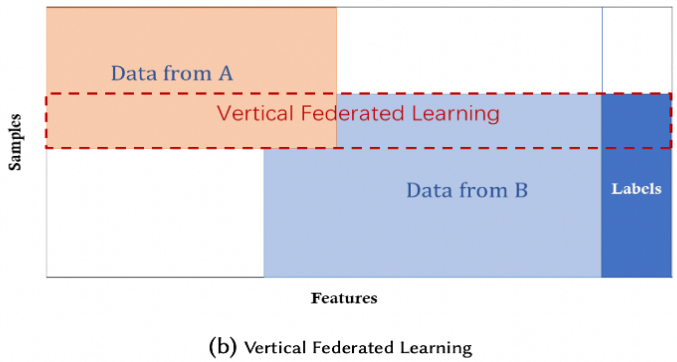
資料來源：Yang, Liu et al. (2019)

DOI:10.6814/NCCU202200065

2. 縱向聯邦學習

縱向聯邦學習採用的數據集之間，特徵不一樣，但是樣本一致。縱向聯邦學習可以表示為

$X_i \neq X_j$，$Y_i \neq Y_j$，$I_i = I_j$，對於任意不同兩家企業 i 和 j，有類似的客戶，但是所記錄的特徵不同。縱向聯邦學習按照特徵劃分數據，增加了數據特徵，各方標籤共享。縱向聯邦學習通常用於跨機構的場景。同樣，多種模型例如線型模型、SecureBoost、神經網路等等都已經在縱向聯邦上實現。



(b) Vertical Federated Learning

圖 4　縱向聯邦式學習

資料來源：Yang, Liu et al. (2019)

3. 遷移式聯邦學習

遷移式聯邦學習的數據集樣本和特徵都不相同。可以表示為：

$X_i \neq X_j$，$Y_i \neq Y_j$，$I_i \neq I_j$，對於任意不同兩家企業 i 和 j，客戶和所記錄的特徵均不相同。遷移式聯邦學習是將兩家客戶，特徵均不相同的公司通過學習某一家公司的特徵分佈，將其信息遷移到另一家公司。例如可以將對於大型企業貸款的模型信息，遷移到數據不足的小型企業融資模型中，以提升小型企業數據不足、不全面等

DOI:10.6814/NCCU202200065

問題。和前兩種模型相比，遷移式聯邦學習的研究還較少，是聯邦學習未來的研究重點。

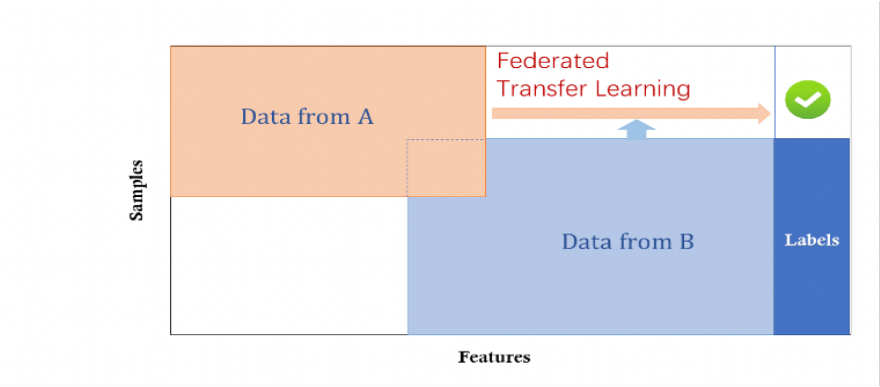


圖 5　遷移式聯邦式學習

資料來源：Yang, Liu et al. (2019)

三、聯邦學習實現用戶隱私保護

聯邦學習是隱私保護的分布式機器學習框架。關於隱私問題，目前在聯邦學習中最常用的是安全多方計算、同態加密、差分隱私。

安全多方計算理論是多組不互相信任的參與方在隱私保護問題和沒有可信第三方的前提下，解決協同計算問題的理論。在無可信第三方的前提下通過數學理論，保證參與計算的各方成員輸入信息不暴露，並得到準確的運算結果。例如品牌希望對其一方數據等，利用第三方數據對其進行標簽增補，完善用戶畫像，強化數據價值。在此過程中，首先先將加密後的品牌數據生成隨機的 ID，隨後將聲稱的隨機 ID 上傳至品牌設置的安全環境中，之後將隨機 ID 發送給第三方數據服務商。服務商在其保險櫃中將匹配上的隨機 ID 添加標簽，最終返回到品牌自己控制的環境中。在此過程中雙方的匹配是基於假名化並加密過後的隨機 ID, 而沒有匹配上的隨機 ID 無法逆向工程再辨識以及使用，保障一方的數據安全。

同態加密是一種對加密進行處理的方法，允許對密文執行數學運算二部解密他們。其他人可以對加密數據進行處理，但是處理過程不會洩露原屬數據，當擁有密鑰的用戶對處理過的數據進行解密，可以得到處理後的結果。

DOI:10.6814/NCCU202200065

在差分隱私中，數據通常是由可信第三方進行加噪音以便保護個人隱私。聯邦學習的差分隱私包含三個步驟：剪裁、聚合、加噪。在服務器收到加密梯度後，並對梯度聚合之後，中心服務器添加高斯噪聲，再進行模型參數的更新。

四、聯邦學習模型介紹

1、 聯邦線型模型

線型模型也包括了邏輯回歸，主要採用了同態加密的思想，對數據和梯度加密。假設存在兩方本地數據 A、B 和中心服務器。

首先由中心服務器向 A、B 發送建模金鑰，A、B 分別初始化參數，並計算梯度，計算完畢後將計算結果加密後發送給中心服務器，並幫助 A、B 更新參數。

2、 聯邦樹模型

Liu, Y. ,et al. (2020)提及了聯邦森林，在建模過程中，每棵樹都實行聯合建模，其結構被存儲在中心服務器及各個數據持有方，但是每個數據持有方僅持有與己方特徵匹配的分散節點資訊，無法獲得來自其他數據持有方的有效資訊，以保障數據的隱私性。 最終整個隨機森林模型的結構被打散存儲，中心服務器中保留完整的結構資訊，節點資訊被分散在各數據持有方。 在使用模型進行預測時，首先獲取本地存儲的節點資訊，然後通過中心節點聯合調用樹結構中其他用戶端的節點資訊。

SecureBoost 可以看作是一種去中心的梯度提升決策樹，通過聯邦學習的方式進行構建樹模型，並交換訓練的權重，通過同態加密保證數據安全。

3、 聯邦支持向量機

V Hartmann,et al.(2019)提出了一種支持向量機（support vector machine，SVM）安全部署在聯邦學習中的方法，主要通過特徵哈希、更新分塊等方式對數據隱私進行保證。

五、FATE 平台構建聯邦學習模型

DOI:10.6814/NCCU202200065

在接下來的章節中，將更加詳細地闡述如何使用目前較流行的平台 FATE 構建橫向聯邦學習。FATE 在組建配置開展聯邦學習項目具有一定的優勢：

(1) 提供安全策略機制，如同態加密，秘密共享等安全計算協議，保證數據隱私安全。

(2) 提供可視化介面 FATEBoard，方便檢查模型和結果。

(3) 支持常用的機器學習算法，包括了特徵工程、模型訓練、模型評估的完整流程。

在此章節中對於下文會使用到的橫向聯邦學習會採用的梯度下降（FederatedSGD）的方式進行迭代優化，下文將簡單闡述：

聯邦學習 server 和 worker 端，worker 擁有本地數據，並進行訓練，而 server 負責匯總各個 worker 節點的訓練結果，並向 worker 發送參數。

假設存在 k 個不同的 worker，第 i 個 worker 擁有的數據量為 n_i，n 為 k 個客戶端數據的總合。因此滿足：

$$\Sigma_{i=1}^{k} \frac{n_i}{n}=1$$

聯邦學習的過程中，每一次迭代會在各個客戶端（worker）進行梯度計算，然 e 輪迭代。

假設在第 t 輪迭代下，sever 端發送參數為w_t，而第 i 個 worker 節點收到後進行訓練，得到損失函數$F_i(w_t)$，並計算梯度g_i。

$$g_i = \nabla F_i(w_t)$$

Worker 將g_i傳輸給 sever 端後，由 sever 進行加權，更新參數w。

$$w_{t+1} \leftarrow w_t - \eta \Sigma_{i=1}^{k} \frac{n_i}{n} g_i$$

η為學習效率。

上述為 FedSGD 的過程。而在聯邦學習中，數據傳輸的通信成本遠高於再本地數據的訓練成本。為了減少 FedSGD 在每次迭代中只進行一次本地訓練就開始傳輸

DOI:10.6814/NCCU202200065

造成的通信成本，因此，對於 FedSGD 改進後採用 FedAvg 進行聯邦優化，過程如下：

在計算出g_i，在本地進行迭代和更新參數，重複下列程序多次：

$$w_{t+1}^{i} \leftarrow w_t - \eta g_i$$

得到新的參數$\widehat{w_{t+1}^{i}}$，並傳送給 sever，而由 sever 端對於所有 worker 端傳輸的結果更新參數。

$$w_{t+1} \leftarrow \Sigma_{i=1}^{k} \frac{n_i}{n} \widehat{w_{t+1}^{i}}$$

DOI:10.6814/NCCU202200065

第三章　實證研究

第一節　數據前處理

由於聯邦學習的訓練所採用的數據來自不同的數據方，而對於不同的數據方，其各自擁有的數據相關性存在差异、數據標量不一以及資料類型不一，因此，多個數據方訓練各自模型的學習速率相差較大，使得模型學習過程中出現較大波動，學習效率較低，導致各機构訓練模型時耗費較大成本。董厶溢（2020）所發表的基於聯邦學習數據處理方法、裝置、設備及介質與流程中說明：

首先先確定客戶端共有的用戶特徵數據、對用戶特徵數據進行特徵編碼處理、得到待處理特徵數據，並獲取基於待處理特徵數據進行處理所得到的模型預測值，之後採用預定義的損失函數對訓練標籤數據和模型預測值進行處理所得到的損失值。若損失值為外部損失值，則將外部損失值以加密管道發送給另一客戶端； 若損失值為內部損失值，則基於內部損失值和第一待訓練模型對應的當前模型參數，確定目標梯度，根據目標梯度對第一待訓練模型進行模型優化，獲取目標預測模型。採用該方法可以增加模型訓練的準確性。

而本文中，數據採用的是同一個數據庫 SEER 的資料，因此，特徵相同，數據標籤和資料類型相當，所以採用傳統的機器學習方法進行數據前處理。

本研究採用數據來自於 SEER 資料庫，選取了 2010 年至 2016 年的肺癌資料，原始資料共 23 萬筆。SEER 資料庫中初始數據中缺失值比例佔比極高，且存在部分特徵重複，為了得到更加適合訓練的數據，本研究首先對部分特徵值進行了刪減。例如可以代表人種（Race）的幾個特徵，分別採用了不同的劃分標準，例如 RACE RECODE (WHITE, BLACK, OTHER)下只分為黑人，白人，其他，和 UNKNOW，而類似的特徵 RACE/ETHNICITY 下一共有 31 個值（具體值見下圖），本研究保留分類更加細緻的特徵 RACE/ETHNICITY。秉持同樣的原則，對特徵進行刪減後，並將有缺失資料的樣本刪除，處理完的樣本一共包括了 13,1336 筆資料，49 個特徵值。

DOI:10.6814/NCCU202200065

Code	Description
01	White
02	Black
03	American Indian, Aleutian, Alaskan Native or Eskimo (includes all indigenous populations of the Western hemisphere)
04	Chinese
05	Japanese
06	Filipino
07	Hawaiian
08	Korean (Effective with 1/1/1988 dx)
10	Vietnamese (Effective with 1/1/1988 dx)
11	Laotian (Effective with 1/1/1988 dx)
12	Hmong (Effective with 1/1/1988 dx)
13	Kampuchean (including Khmer and Cambodian) (Effective with 1/1/1988 dx)
14	Thai (Effective with 1/1/1994 dx)
15	Asian Indian or Pakistani, NOS (Effective with 1/1/1988 dx)
16	Asian Indian (Effective with 1/1/2010 dx)
17	Pakistani (Effective with 1/1/2010 dx)
20	Micronesian, NOS (Effective with 1/1/1991)
21	Chamorran (Effective with 1/1/1991 dx)
22	Guamanian, NOS (Effective with 1/1/1991 dx)
25	Polynesian, NOS (Effective with 1/1/1991 dx)
26	Tahitian (Effective with 1/1/1991 dx)
27	Samoan (Effective with 1/1/1991 dx)
28	Tongan (Effective with 1/1/1991 dx)
30	Melanesian, NOS (Effective with 1/1/1991 dx)
31	Fiji Islander (Effective with 1/1/1991 dx)
32	New Guinean (Effective with 1/1/1991 dx)
96	Other Asian, including Asian, NOS and Oriental, NOS (Effective with 1/1/1991 dx)
97	Pacific Islander, NOS (Effective with 1/1/1991 dx)
98	Other
99	Unknown

圖 6　　研究採用特徵之一 RACE/ETHNICITY 的值

資料來源：RESEARCH PLUS DATA DESCRIPTION

其中，特徵包括了人種，年齡，性別，診斷時間，癌症治療方式：是否進行手術，放療（Radiation recode），化療（Chemotherapy recode (yes, no/unk)）等，癌症發病特性：癌症等級（Grade）、分化部位(Primary Site)，腫瘤大小（CS tumor size (2004-2015)），發病部位單雙側（Laterality），是否轉移：包括骨轉移（SEER Combined Mets at DX-bone (2010+)），腦轉移（SEER Combined Mets at DX-brain (2010+)），肝轉移（SEER Combined Mets at DX-liver (2010+)）等等資料，共 49 個特徵變數進行分析。由於資料中存在 string 格式的數據，這裏將資料採用了 Label Encoder 的方式對於有順序的 string 資料進行格式轉換，例如癌症等級（Grade）分為 I，II，III 級，按照分化程度進行排序；腫瘤大小（CS tumor size (2004-2015)）按照腫瘤長度進行排序。而對於癌症是否轉移，發病部位單雙側等資料則採用 OneHotEncoder 進行轉換。

DOI:10.6814/NCCU202200065

圖 3-1 是對肺癌存活月份的分佈圖，圖中顯示肺癌的生存率隨著月份的增加而減少，最高存活月數為 84 個月。由於肺癌的死亡率極高，且治療困難，因此可以看到肺癌的存活時間較短。而根據資料，存活超過兩年的人數為 4,6870 人，佔全部人數的 35.69%。

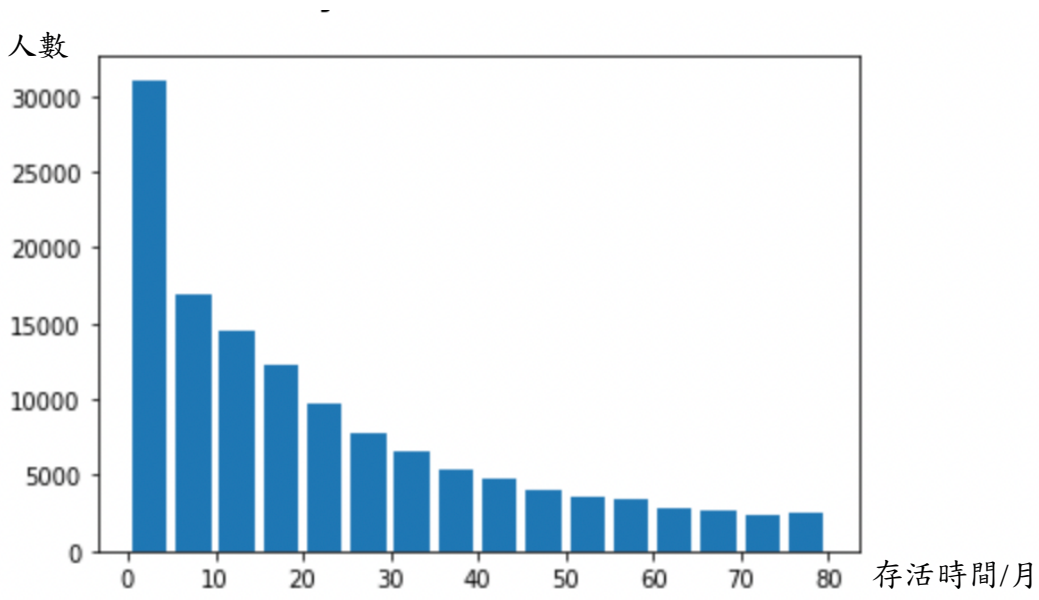


圖 7　肺癌存活月數分佈

資料來源：本研究整理

表 3-1 則展示了其中部分的特徵的敘述統計量。

表 3　變數敘述統計量

變數名稱	平均數	變異數	最小值	25%分位數	50%分位數	75%分位數	最大值
Year of diagnosis 治療年份	2012.48	1.704271	2010	2011	2012	2014	2015

DOI:10.6814/NCCU202200065

變數名稱	平均數	變異數	最小值	25%分位數	50%分位數	75%分位數	最大值
Age at diagnosis 開始治療的歲數	69.08222	10.40461	2	62	70	77	103
Sex 性別	0.52	0.50	0	0	1	1	1
Race/ethnicity 人種	22.39	8.12	0	26	26	26	26
Primary Site 腫瘤部位	342.18	2.10	340	341	341	343	349
Histologic Type ICD-O-3 腫瘤型態編碼	8137.54	116.72	8000	8070	8140	8140	8940
Grade 分化等級	2.40	0.69	1	2	3	3	3
Laterality 單雙側	3.31	1.98	0	1	5	5	5

資料來源：本研究整理

為了預測存活率，根據之前的大部分涉及癌症存活率預測的研究中，大部分採用了 2 年期和 5 年期的存活時間，因此本研究中也採取同樣的標籤，將預測的 y 值設定為 2 年存活率和 5 年存活率，隨機抽取其中的 80%為訓練集，20%為預測集，通過機器學習的方式分別預測了值。2 年內肺癌存活率的平均值是 35.69%,而 5 年內存活率僅為 8.02%。

第二節　傳統機器學習實證研究

本研究建立了 6 種不同的機器學習方式，包括了決策樹、隨機森林、XGboost、邏輯回歸、支持向量機和 knn 的方法，經過調整參數後得到預測結果。2 年內存活率的研究結果顯示 XGboost 的預測結果最好，正確率達到了 82%。而支持向量機的方式結果較差，只有 71%。

DOI:10.6814/NCCU202200065

研究結果整理如下表：

表 4　2 年內存活率預測結果

模型		accuracy	precision	recall	F1
決策樹	Train	0.82458	0.7448621	0.75977	0.75224
	Test	0.82062	0.73598	0.75231	0.74405
隨機森林	Train	0.85406	0.78565	0.80209	0.79378
	Test	0.82439	0.74038	0.75823	0.7492
XGboost	Train	0.85889	0.8091	0.79881	0.80392
	Test	0.82503	0.75801	0.75059	0.75428
邏輯回歸	Train	0.75416	0.65302	0.65718	0.65509
	Test	0.75598	0.65313	0.65637	0.65475
knn	Train	0.81184	0.7122	0.74913	0.7302
	Test	0.73066	0.60058	0.62468	0.61239
支持向量機	Train	0.85406	0.78565	0.80209	0.79378
	Test	0.82439	0.74038	0.75823	0.7492

資料來源：本研究整理

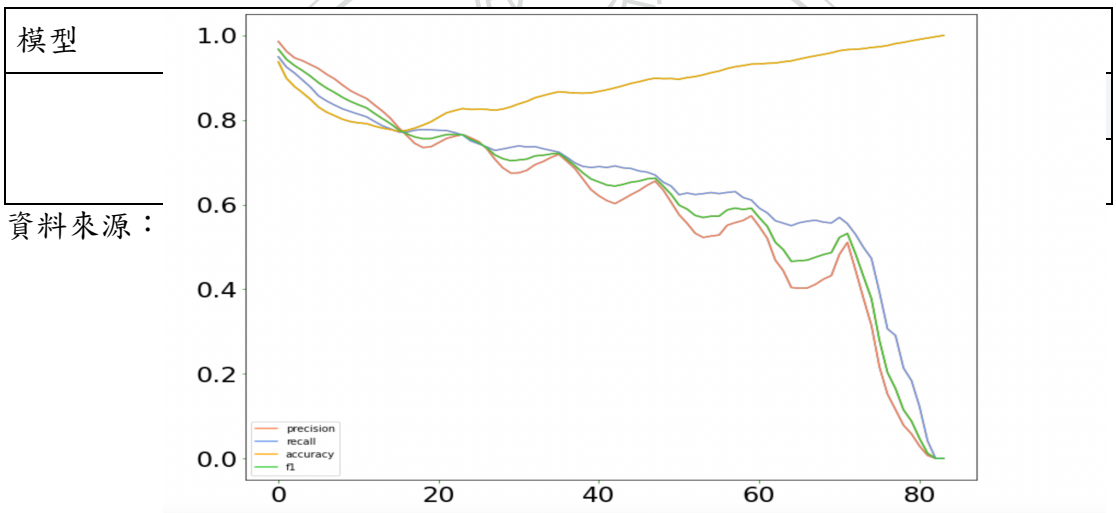
如上表所示，隨機森林，決策樹，XGboost 的預測結果相近，但是其他預測結果較為不盡人意。而對於 5 年內的研究結果，預測結果較為糟糕，雖然正確率很高，但是 recall 和 precision 都普遍較低。在這裡只列出隨機森林的結果。

DOI:10.6814/NCCU202200065

表 5　五年內存活率預測結果

可以看到上述結果中，accuracy 達到了 0.93。之所以產生上述結果，是由於資料偏誤所產生，肺癌 5 年內的存活率只有 8%，因此在預測的時候，會造成準確率很高，但是 recall 和 precision 較低的現象。

以不同的存活月數為界而判定出的存活率有較大的差異，正如上文所述，當設定的存活時間越長，會導致存活概率大幅度小將，資料偏誤會愈發嚴重。為了觀測由於資料的偏誤造成的影響，因此，我們對於存活月數進行了分析。通過 1-84 個月的存活率分別進行了預測，檢驗數據的偏誤對結果的影響。圖 3-2 為以不同的存活月數為界判定存活率下，隨機森林模型的測試集結果。



資料來源：

圖 8　不同存活月數的檢驗結果

資料來源：本研究整理

結果顯示，存活月數低於 19 個月為界劃分資料，accuracy 隨著月數的增加而減少，但是超過 19 個月之後，accuracy 則隨之增加。而 recall，precision 和 F1 則隨劃分界線的月份的增加一直存在下降的趨勢，且變動趨勢一致。當劃分界線超過 50 個月時，precision 低於 60%，預測結果逐漸出現 accuracy 很高，但是其他模型評估指標很低的情況。Ganganwar, V. (2012)中提及：如果一個數据集包含的一個類的樣本比其他類的樣本多，則稱為不平衡數据集。 當至少一個類別的數據量過少，而其他類占多數時，數据集是不平衡的。 在這種情況下，分類器在大多數類上具有良好的準確性，但在其他的指標上，由於較大多數類對傳統訓練標準的影響，具有非常差的準確性。因此，由於五年存活率只佔了 8%，從而導致了上述結果。因此，本文剔除了五年存活率的指標，下文僅採用 2 年存活率進行訓練。

DOI:10.6814/NCCU202200065

第三節　聯邦式學習實證研究

本節希望可以和傳統機器學習進行對比，研究是否採用聯邦學習再考慮數據安全，並且在數據孤島的情況下是否還能夠產生和傳統機器學習類似的學習效果。該章節對於上文建立模型所採用的數據進行劃分，採用 SecureBoost 驗證。

由於實證採用的數據特徵都一致，選取了由微眾銀行所建立的 Fedrated AI Enabler 平台（後文簡稱 FATE）進行聯邦橫向學習。在此過程中，由於 SEER 資料庫的權限問題並不能取得每筆資料所來源的醫院。所以我們根據 FATE 平台所提供的實際案例所採用的方法對數據集進行最簡單的劃分。對於整理出的 13,1336 筆資料，同樣隨機抽取 80%作為訓練集（共 10,5068 筆數據），20%作為預測集（共 2,6268 筆數據）。而將訓練集的部分隨機抽出 5,0000 筆作為 host，即為數據提供方，而剩下 5,5068 筆資料作為 guest，即為數據應用方。預測集則為 guest 和 host 二方共用。

訓練過程如下（可參考 FATE 官方文件）：

(1)數據轉換輸入：將原本 csv 格式轉換為 FATE 所要求的 Dtable 格式。

(2)進行模型訓練：構建模型，採用 FATE 自訂的領域特定語言（DSL）對於各種模塊組織起來，組建各種算法並且進行參數配置。模塊包括了數據讀寫（data_io），特徵工程（feature-engineering），回歸（regression），分類（classification）。

(3)評估模型：修改模型訓練中的配置，檢查模型效果，由 FATEBoard 進行結果可視化。

表 6 展示了進行 SecureBoost 模型後所產生的結果，訓練集和測試集的評估相差不大。測試集中，準確度達到了 0.863967，而 precision 為 0.704097。

表 6　SecureBoost 結果

模型		Auccuracy	Precision	Recall	F1
SecureBoost	Train	0.870403	0.696873	0.792348	0.74155
	Predict	0.863967	0.704097	0.797848	0.74805

資料來源：本研究整理

DOI:10.6814/NCCU202200065

Fateboard 還提供了 ROC 曲線如下圖所示。

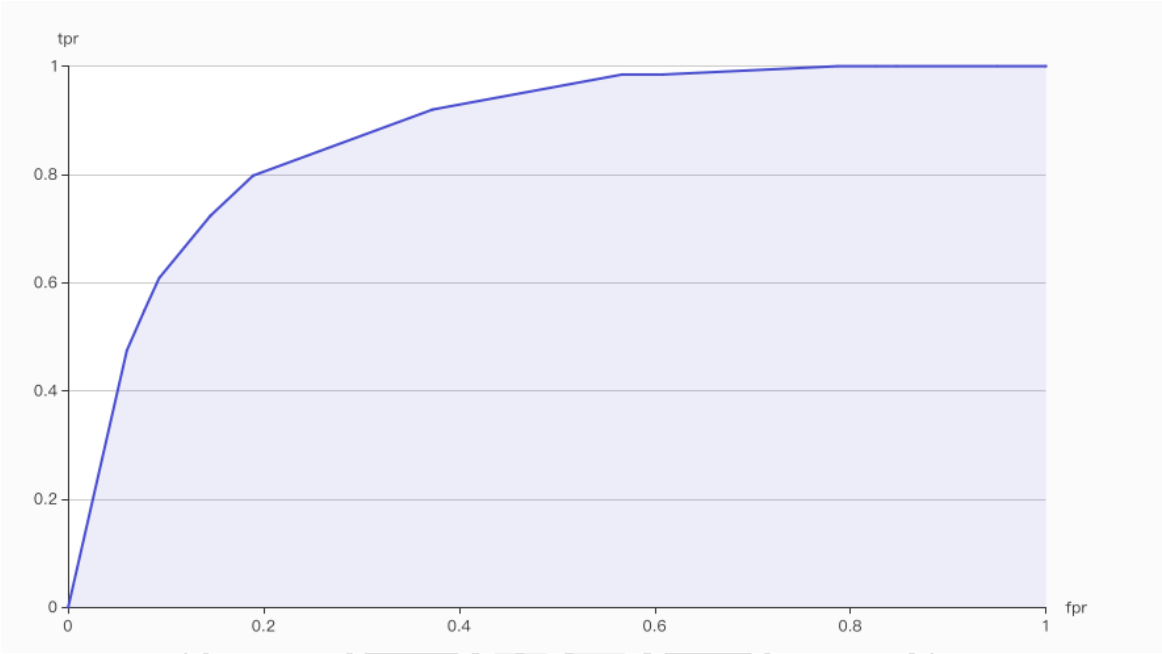


圖 9　SecureBoost 模型 ROC 曲線

資料來源：FATEBoard

該章節主要描述了在 FATE 平台上如何構建 SecureBoost 模型，並給出了訓練結果。在上文中已經提及，由於 5 年期存活率會有很大的偏誤，因此該章節僅採用了 2 年期肺癌存活率作為標籤。在下節中，我們將對傳統機器學習模型和聯邦學習模型結果進行比較。

第四節　聯邦學習與傳統機器學習效果比較

該節對比了二、三節中的模型評估結果，以此來判定聯邦學習是否可以在醫院數據不能相互傳送的基礎上得到和傳統機器學習相同的效果。整理後表格如下：

表 7　聯邦學習和傳統機器學習效果比較

模型		Auccuracy	Precision	Recall	F1

DOI:10.6814/NCCU202200065

Secureoost	0.870403	0.696873	0.792348	0.74155	0.870403
	0.863967	0.704097	0.797848	0.74805	0.863967
XGboost	Train	0.85889	0.8091	0.79881	0.80392
	Test	0.82503	0.75801	0.75059	0.75428

資料來源：本研究整理

為了表格簡潔起見，我們只將機器學習中效果最好的 XGboost 進行比較。在本研究中所假設的兩家不能互相傳輸數據的醫院場景下聯邦學習的效果並沒有減弱。這說明，聯邦學習在一定程度上可以解決數據孤島的問題，幫助保險公司獲取更多的信息資料，以更加準確的預測癌症險的出險和理賠。除了癌症險之外，保險公司也可以將聯邦學習應用在更多的險種上，面對越來越嚴格的數據保護政策，也為保險公司提供了更多的可能性。

DOI:10.6814/NCCU202200065

第四章　結論

在數據保護愈發嚴格的情勢下，保險公司為了預測癌症險出險理賠的過程也愈發困難，如何在保護本地的數據的同時，獲得更多的數據樣本進行分析對於保險公司不可或缺。本研究展示了一種解決數據孤島問題的方式，通過 FATE 聯邦學習平台，展示了肺癌存活率的預測過程，並且對比了聯邦學習和傳統機器學習的訓練效果。結果發現聯邦學習並不會降低模型訓練的效果，在一定程度上為保險公司應對數據保護的問題提供了解決的方向。

但本研究在一定程度上有所缺失，還可以有補足的方向:

(1) 嘗試採用更加多元化的聯邦學習模型進行探索，以得到更精確的訓練效果。採用縱向聯邦學習，構建保險公司和醫院或者其他企業的模型。

(2) 聯邦學習的分群模式過於簡單，可能和現實情況有所不符。如何更加有效地進行分群，以達到和現實中類似的效果。

(3) SEER 資料庫中有部分特徵由於權限問題無法獲取，例如：病人是否購買保險，婚姻狀況等等，可能對於建立模型有更大的幫助。

希望在未來可以更加深入地對於聯邦學習進行探討，更加有效的解決目前數據孤島的難題。目前的研究還沒有辦法完全解決在聯邦學習的過程中遭到惡意攻擊的問題，期待未來可以真正通過聯邦學習的方式，將數據安全並有效地進行應用。

本文只採用了橫向聯邦學習的模型，而在保險公司實際的建構模型中，與其他企業例如醫院，銀行等等進行合作，獲取數據更多的是通過縱向聯邦學習進行構建，未來希望可以找到合適的資料數據進行分析。

DOI:10.6814/NCCU202200065

參考文獻

中文部分：

1. 周脈耕,王黎君,黃正京,楊功煥.2002.人口老化及危險因素改變對肺癌死亡率的影響[J]. 中國衛生統計.
2. 李媛秋, 劉劍君，麼鴻雁 .(2019). "肺癌發病和死亡流行情況与人類發展指數展的關係分析." 中國腫瘤 28(9): 646-650.
3. 楊強，黃安埠，劉洋，陳田健.(2021).聯邦學習實戰.
4. 馬立偉, 曾強, 呂秋平, 范成燁， & 程鵬. (2015). 大數据癌症風險預測系统. 世界复合醫學(1), 5.
5. 衛生福利部中央健康保健署.(2019). 醫療支出費用.
6. 潘憶文(I-Wen Pan), 簡君儒(Chun-Ru Chien), & 施雅真(Ya-Chen Tina Shih). (2012). 美國癌症登記及老人醫療保險資料庫之發展與應用－論台灣癌症登記與健康保險聯結資料庫之可行性. 台灣公共衛生雜誌, 31(4), 299-310.
7. Thomas Wetter. (2006). 運用三種資料探勘方法預測子宮頸癌存活情形之比較. 台灣家庭醫學雜誌, 16(3), 192-203.
8. 胡麗霞, 江長思, 羅燕, 梅東東, 龔靜山, & 馬捷. (2019). 基于機器學習的放射組學預測肥細小肺癌 egfr 基因突變. 醫學影像學雜誌, 29(7), 4.
9. 財團法人保險事業發展中心.(2019). 108 年人壽保險業務統計年報.
10. 王健宗、孔令煒、黃章成、陳霖捷、劉懿、何安珣、肖京.（2020）. 聯邦學習算法綜述. 大數據，6（6），19.
11. 董亾溢.（2020）. 基於聯邦學習數據處理方法、裝置、設備及介質與流程 .

英文部分：

1. Yang, Q. , Y Liu, Y Cheng, Y Kang, & Yu, H. . (2019). Federated Learning. Morgan & Claypool.

DOI:10.6814/NCCU202200065

2. Yang, Q. , Liu, Y. , Chen, T. , & Tong, Y. . (2019). Federated machine learning: concept and applications. ACM Transactions on Intelligent Systems and Technology, 10(2), 1-19.
3. Liu, Y. , Liu, Y. , Liu, Z. , Zhang, J. , Meng, C. , & Zheng, Y. . Federated forest. IEEE Transactions on Big Data, PP(99), 1-1.
4. Yang K , Jiang T , Shi Y , et al. Federated Learning via Over-the-Air Computation[J]. 2018.
5. V Hartmann, Modi, K. , Pujol, J. M. , & West, R. . (2019). Privacy-preserving classification with secret vector machines.
6. WILD, C. P., E. WEIDERPASS and B. W. STEWART (2020). Cancer Report :Cancer research for cancer prevention.
7. Huang, L. , & Liu, D. . (2019). Patient clustering improves efficiency of federated machine learning to predict mortality and hospital stay time using distributed electronic medical records. Journal of Biomedical Informatics, 99, 103291-.
8. Ferlay, Shin, Bray, & Mathers. (2010). Globocan 2008, cancer incidence and mortality worldwide: iarc cancerbase no. 10. International Journal of Cancer Journal International Du Cancer, 136(5), E359 – E386.
9. Xia, Y. , Yang, D. , Li, W. , Myronenko, A. , Xu, D. , & Obinata, H. , et al. (2021). Auto-fedavg: learnable federated averaging for multi-institutional medical image segmentation.
10. Rehak, D. R. , Dodds, P. , & Lannom, L. . (2005). A model and infrastructure for federated learning content repositories.
11. Mcmahan, H. B. , Moore, E. , D Ramage, Hampson, S. , & Arcas, B. . (2016). Communication-efficient learning of deep networks from decentralized data.
12. Peter Kairouz, H.Brendan McMahan, Brendan Avent, & et al. (2019). Advances and open problems in federated learning.
13. He, H. , & Garcia, E. A. . (2009). Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering, 21(9), 1263-1284.

DOI:10.6814/NCCU202200065

14. Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering, 2(4), 42-47.

DOI:10.6814/NCCU202200065