ISSN 1330-3651 (Print), ISSN 1848-6339 (Online) https://doi.org/10.17559/TV-20240601001729
Original scientific paper

VFed-PU: Identifying Containers with Potential to be Shipped by Rail from Ports with Privacy Protection

Lei HUANG*, Deyou JIANG, Xiong ZHANG, Ying WANG, Tianyang BAI

Abstract: Facing challenges in the global container shipping market and strict data protection laws like GDPR and CCPA/CCPR, the sea-rail intermodal transportation sector urgently needs better freight demand forecasting. This study develops a micro-level transportation demand forecasting model tailored for the sea-rail intermodal sector, emphasizing data privacy and accurate prediction of port demand for container shipments by rail, which is crucial for effective railway planning and marketing. We introduce a novel framework, VFed-PU, which combines Vertical Federated Learning with Positive and Unlabeled Learning. This model tackles issues such as limited labeled data, data imbalance, and selection bias using a new method called ImbalancednnPUSB. VFed-PU ensures data privacy by transferring only data representations rather than the original data during model training, safeguarding sensitive information among different parties. Extensive experiments demonstrate that VFed-PU outperforms state-of-the-art algorithms in predicting port demand for container shipments, achieving a recall rate of approximately 90%. This framework not only enhances prediction accuracy and preserves data privacy but also supports strategic railway planning and marketing efforts. The study highlights the importance of data privacy in transportation planning, especially under stringent data protection regulations, and contributes significantly to the field by addressing both forecasting performance and privacy concerns.

Keywords: freight demand forecasting; positive and unlabeled learning; sea-rail intermodal transportation; vertical federated learning

1 INTRODUCTION

The global container shipping market is experiencing a significant downturn driven by economic volatility and trade deceleration [1]. In this market downturn, multimodal transportation, meaning using multiple types of transport together to deliver goods, has become particularly pertinent due to its congestion relief, environmental friendliness, high efficiency, and flexibility [2]. The intermodal freight transportation market is anticipated to achieve a compound annual growth rate (CAGR) of 8.27% from 2023 to 2028 [3]. Meanwhile, customers are expecting more shipping options with higher delivery speeds, higher reliability, lower shipping prices, more flexible destination options, easy returns, and simple tracking [4].

Facing such challenges and market needs, accurate freight demand forecasting is extremely important for intermodal carriers to plan shipments, understand market trends, and improve their market competitiveness. Better data understanding and appropriate models are the keys to fulfilling these needs while planning and managing the supply chain process [4]. However, with the implementation of the General Data Protection Regulation (GDPR) in the EU and CCPA/CCPR in the USA, as well as the general public's growing awareness of privacy protection, it has become increasingly difficult for intermodal carriers to access extensive customer and market data from multiple sources, especially in sea-rail intermodal container transportation.

Data protection laws have significantly challenged traditional approaches to predict freight demands. In previous research, freight demand is usually modeled using regression and time series techniques [5], input-output models, and machine learning methods [6]. Among these methods, the original data plays a critical role in data analysis and feature extraction. However, with the increasing enforcement of data protection laws and the rising concerns over privacy, traditional models may no longer be viable, as they often require access to sensitive and proprietary data. This highlights the need for a new framework that not only predicts freight demand with high precision but also complies with privacy regulations. While some research focuses on predicting future freight volumes at the macro level using macroeconomic indicators and time series data, such approaches are insufficient for supporting specific operational decisions, like determining the need for railway transport at the port level [7].

This study aims to fill this gap by identifying and predicting the transportation demands for standard shipping containers in sea-rail intermodal transportation at the micro level and, more importantly, doing so with a focus on data privacy protection for data owners. The process of identifying containers at the port that may require railway transport relies on the collaborative analysis of multisource data, including, but not limited to, real-time freight information from ports, railway transportation data, shipper demands, and historical transport records. Therefore, data privacy protection is a key concern during data analysis and model training to identify containers currently at a port for which the port will require railway transport; we refer to this identification process as mining potential railway containers.

To solve this problem, this study proposes an innovative framework, VFed-PU, for discovering freight container demand at the port while also preserving ports' privacy during model training and implementation. This framework integrates Vertical Federated Learning with Positive and Unlabeled Learning, and the proposed algorithm *ImbalancednnPUSB* is specifically designed to address issues of data imbalance and selection bias in this research context. This approach can assist railways in identifying the transportation container demand at ports in advance, enabling them to carry out marketing activities more effectively. By doing so, railways can increase the volume of container sea-rail intermodal transportation, thereby facilitating a transportation shift from roads to railways. This shift is crucial for achieving sustainability and efficiency in logistics as, compared with railway transport, road transport often suffers from higher congestion, higher carbon emissions, and higher susceptibility to weather and traffic conditions.

The contributions of this study are as follows. First, different from previous research predominantly focusing on horizontal federated learning, we develop a groundbreaking federated learning framework, VFed-PU, by integrating vertical federated learning with positive and unlabeled learning. Second, we propose the ImbalancednnPUSB approach based on existing PU learning methods and specifically tailor it to address the complexities of real-world data scenarios. Third, through field studies, we show that identifying containers with potential to be shipped by rail from ports is an important practical need for railways, and we clearly define this problem. We propose a framework to efficiently address this issue and safeguard data privacy at the same time. Fourth, we find that for potential railway container transport without price discounts, the differences in transportation duration are typically between 40 to 100 hours. Railways face more intense competition with roads for close destinations, as indicated by a higher cost difference, especially for stations within the 500-1000 km range, and the time differential percentages show no significant variation between short- and long-distance travel to stations. Finally, our approach can provide detailed guidance for the transport planning and pricing policy of each railway station.

2 LITERATURE REVIEW

2.1 Freight Demand Forecasting

In the domain of freight demand forecasting, the literature adopts various types of models, including time series models, regression models, machine learning models, econometric models, and simulation models. These can be used in the problem of predicting port demands for railways to ship containers.

Time series models, such as ARIMA, mainly try to leverage the trends of historical data. In contrast, regression models usually use variables like economic indicators to predict freight demand. Yang and Yu [8] constructed four multivariate statistical regression methods - ordinary least squares regression (OLSR), principal component regression (PCR), partial least squares regression (PLSR), and a modified partial least squares regression (MPLSR) - to predict Railway Freight Volume (RFV) and compared their performance on a practical dataset. Khan and Khan [9] utilized standard multivariate time series methods, such as the Johansen co-integration and error correction model, to obtain short-run and long-run elasticities of rail freight transport demand in Pakistan. Other time series methods, like LSTM [10], have also been used in this context.

Machine learning models, such as Random Forests and Neural Networks, can be trained using complex datasets. Salais-Fierro and Martínez [11] demonstrated the superior accuracy of Artificial Neural Networks (ANNs) over traditional statistical methods in forecasting freight transport demand using historical data from a transportation management system.

Econometric models are also used for freight prediction. C. Lu et al. [12] constructed an input-output model to analyze and forecast freight demand by examining the impact of economic growth, industrial structure changes, and complete consumption coefficients. They found that economic expansion significantly contributes to transportation value increases and that consumption coefficient shifts drive freight traffic growth.

Simulation models, such as agent-based models, can assess system behaviour in various scenarios. Nuzzolo and Comi [13] presented an urban freight demand forecasting model integrating quantity, delivery, and vehicle-based subsystems and simulating the impacts of different logistics measures.

Hybrid models combine features from different models to potentially achieve better performance. Feng et al. [14] proposed an ensemble model composed of SARIMA and a deep belief network to predict railway freight volume. Wang et al. [15] proposed the GCA-GA-GNN model, combining grey correlation analysis, genetic algorithms, and grey neural networks, to accurately predict railway freight volumes.

Traditional freight demand models mainly focus on macro-level forecasting using aggregate data, which overlooks the finer details of individual container transportation. These models fail to address critical aspects like real-time demand, container-specific transit times, and destination restrictions. As a result, they often lead to suboptimal resource allocation. In contrast, a micro-level approach, focusing on predicting transportation demand for individual containers, allows for more precise planning and better resource utilization. This gap in traditional models emphasizes the need for micro-level analysis, which this study aims to address for sea-rail intermodal carriers.

2.2 Vertical Federated Learning

Vertical Federated Learning (VFL) refers to a collaborative model training approach on a dataset with features distributed across multiple parties, with one active party holding sensitive label information and the others serving as passive parties [16]. The active party maintains a global module, which can be either a static aggregation function in aggVFL architecture or a trainable model in splitVFL architecture [17].

In aggVFL, each participant contributes to the model with their local sub-models, and the collective output is synthesized via a fixed global aggregation function. In the literature, various tree-based models, such as SecureBoost [18], SecureGBM [19], and VPRF [20], have been proposed. Homomorphic encryption (HE) and other secure multiparty computation techniques are typically employed to enhance the model's security. On the contrary, SplitVFL employs a dynamic, trainable global model that evolves to improve aggregation based on the learning process. SplitVFL is based on the concept of vertical split learning, mainly NN-based [21]. It ensures that participants contribute to the model training without accessing sensitive label information, as the server alone retains the trainable global model, significantly increasing label security [22].

These VFL architectures provide solutions for training models on vertically partitioned data while preserving data privacy and security. They have been applied in various fields, including transportation [23, 24], finance, healthcare, and wireless communication. In particular, the splitting techniques in SplitVFL enhance the privacy guarantees and reduce the risk of label leakage in VFL scenarios.

2.3 Positive and Unlabeled Learning

Positive-Unlabeled (PU) learning represents a variant of the classical binary classification problem. In this scenario, the training data consists solely of positive samples, while the test data is composed of both unlabeled positives and unlabeled negatives. Recent advances in PU learning can be broadly categorized into two sub-streams.

The first sub-stream focuses on modeling the label noise, unbiasedly estimating risk, and adapting existing supervised classification methods to the PU learning setting. Notable contributions in this area include works such as uPU [25] and nnPU [26]. uPU demonstrates that a cost-sensitive classifier can be used by reforming the risk in the original classification with a known class prior. nnPU addresses the overfitting tendency of complex models in uPU by imposing a nonnegative constraint on the objective function, leading to better generalization ability. These methods usually work under the Selected Completely at Random (SCAR) assumption, where the probability for a positive instance to be labeled is constant and independent of the covariates.

A critical limitation of this approach, however, is that it often overlooks the imbalanced nature of the data, which can result in biased estimates of the risk function. As a result, other studies have introduced more nuanced assumptions to better account for real-world complexities. For example, a new assumption called Selected at Random (SAR) has been proposed [27], where the labeling probability for positive instances is influenced by their covariates, addressing the biases inherent in label selection processes. Additionally, methods designed to address imbalanced data in PU settings, such as cost-sensitive learning and re-sampling techniques, have shown promise in reducing model bias and improving classifier accuracy [28]. Furthermore, biased negative data, common in practical settings, has been tackled through adaptations of standard PU learning methods [29], ensuring more accurate negative instance identification and less distortion in the model's decision boundary.

The second sub-stream of recent methodologies addresses PU learning as a two-step process: (1) identify reliable negative instances among the unlabeled observations, and (2) use a standard supervised or semi-supervised classification method to build the PU classifier. For instance, the Self-PU method [30] leverages self-paced learning to gradually update the base model with newly learned knowledge. Another innovative strategy proposed by C. Xu et al. [31] involves splitting the unlabeled dataset with an early-stop strategy. These methods highlight the importance of strategic sample selection and iterative learning in improving the performance of PU learning models.

3 THE PROPOSED FRAMEWORK

3.1 Case and Problem Formulation

Through investigation in ports and railway yards, we observed that the current transportation of containers in sea-rail intermodal transport highly relies on ports putting forward railway car requests to railways according to shippers' needs. This has led to scenarios where ports, to ensure their transportation needs are met, excessively apply for cars, resulting in railway cars being sent to the port but subsequently exiting the port without being full. Consequently, the railways struggle to ascertain the actual requirements of ports. The information asymmetry between railways and ports has resulted in a bargaining process under supply and demand fluctuations. However, by identifying the containers it might be asked to transport, a railway can more accurately understand the port's actual demand, allowing for more efficient and timely allocation of resources and thereby reducing the wastage caused by empty cars. In such a context, identifying potential containers from ports is a crucial step for railways to improve the efficiency of freight transportation. In this study, potential containers refer to containers for which the port might request rail transportation but which have historically been delivered via transportation modes other than railways.

In this study, we mine port information to find containers with potential for railway delivery. We consider containers stored in the port's yard for two reasons. First, the data associated with these containers is usually very comprehensive and can serve as a good dataset for our analysis. Second, the transportation mode for most of these containers is often undecided, making them ideal candidates to potentially convert to rail transport. This identification process is particularly valuable for sea-rail intermodal container transportation as it can help container scheduling at ports and the subsequent train scheduling at railway stations.

3.2 Sample Alignment and Feature Augmentation for Ports and Railways

In our research context, neither party, whether a port or railway, has all the data attributes in data samples. For instance, for containers transported via roads, railway carriers do not know the corresponding railway distance, freight, and transportation time features (these data are available only after such containers have been transported by rail). Similarly, for containers transported via railways, ports do not know the corresponding road distance, freight, and transportation time features (these data are available only for containers previously transported by road). Due to data security and privacy issues, such data attributes cannot be transferred among ports and railway carriers. However, missing any type of such data attributes would significantly hinder model training for potential container identification. Therefore, in our proposed framework, data owners preprocess their data attributes for sample alignment and feature supplementation. Note that we do not physically integrate data from multiple sources into one dataset during data sample alignment and feature supplementation. We only align the features across datasets to ensure consistency in the modeling process while maintaining the privacy of each data source.

The process of sample alignment and feature supplementation is shown in detail in Fig. 1. On the one hand, the port cleans and integrates historical container basic information, shipping schedule data, storage data, and operational data for road and railway transportation. The port then fuses the data into the storage container dataset. Then, data samples not satisfying the criteria are excluded. This selection is according to goods category,

railway station restrictions, time interval, and destination restrictions, as mentioned earlier. On the other hand, the railway carrier cleans and integrates daily demand, waybill, waybill container data, and trajectory data of historical transported containers, and fuses them into the transported container data, which can be aligned with the dataset formed by the port.

After sample alignment, the data of containers historically transported by road cannot be aligned, so feature augmentation is needed, as illustrated in Fig. 1. For railways, the destination of these samples could be used to supplement the model to compensate for the missing railway transportation features. Similarly, the port can supplement its model by predicting the missing road transportation features based on the aligned railway transportation samples.

Ultimately, we align the feature data formed by the port and the railways. This thorough alignment and supplementation process ensures a complete set of features necessary for accurate prediction and identification of potential containers in this research context. Please note that this data alignment and feature supplementation is just for data preparation, not physical data integration.

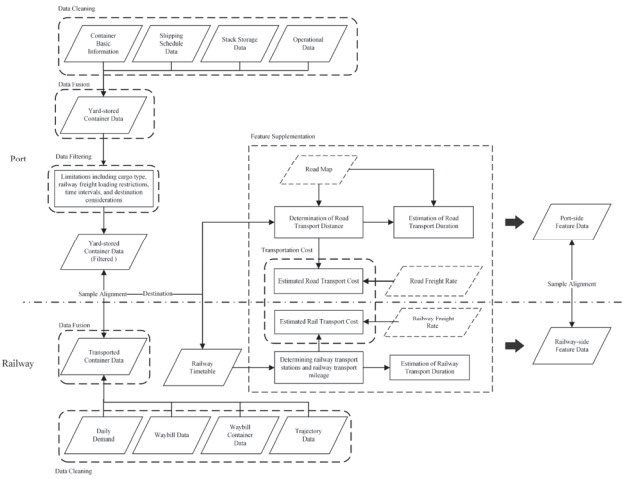

Figure 1 The process of sample alignment and feature supplementation

3.3 VFed-PU Framework

In this context, where the port and railway parties each hold different features of the training samples, a Vertical Federated Learning model is necessary. Containers that have been previously transported by railway can be labeled as positive samples. However, containers transported by road cannot be simply labeled as negative samples since some of them possess characteristics suitable for railway transportation and are possibly the recognition target of our model. This scenario aligns with the PU learning method.

We proposed a new federated learning framework: Vertical Federated Positive and Unlabeled Learning (VFed-PU). This framework includes encrypted sample alignment, class prior calculation, model training, model evaluation, and model explanation, as shown in Fig. 2.

Encrypted sample alignment: This process can be achieved through Private Set Intersection (PSI). It aims to discover common elements across multiple data sources while safeguarding the data privacy of each participant. There is a plethora of specific PSI implementation strategies tailored to various scenarios. In this work, we adopt a PSI implementation based on a Diffie-Hellman key exchange that uses Bloom filter compression to reduce the communication complexity [32]. The computational complexity of the encrypted sample alignment is primarily composed of $O(\log n)$ for the Diffie-Hellman key exchange and $O(m+k)$ for Bloom filter compression and set intersection, where n is the dataset size, m is the Bloom filter size, and k is the number of hash functions used.

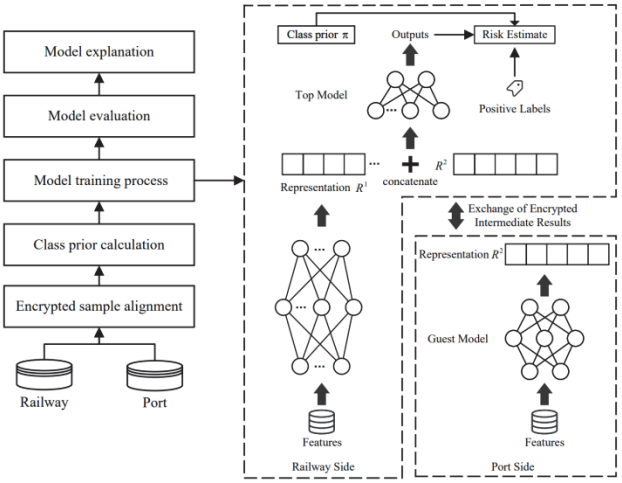

Figure 2 VFed-PU Framework

Class Prior Distribution Calculation: Before model training, it is essential to estimate the class prior distributions, which are utilized for risk estimation in PU learning. We use π to denote the class prior for the positive label. We use the KM2 method for this class prior distribution calculation [33]. To prevent data leakage in scenarios involving vertical partitioning of features, the dot product matrix and squared norms are calculated locally by the data holder. The pseudocode of class prior distribution calculation in our framework is shown in Algorithm 1.

Algorithm 1: KM2 with Feature Vertical Splitting

Input: Feature Matrix M_1, Feature Matrix M_2, and Sample size N

Output: The estimated class prior π

1. Compute dot product matrix $A = M_1 M_1^T + M_2 M_2^T$
2. Compute squared norms vector
$$S = \sum (M_1 \cdot M_1)^T + \sum (M_2 \cdot M_2)^T$$
3. Compute distance squared matrix
$$D = I_N \otimes S + S^T \otimes I_N - 2A$$
4. Use the median of D as the initial kernel width and generate a logarithmic space sequence around it as candidate *kernel_widths*
5. for σ in *kernel_widths*:
   Compute the Radial Basis Function (RBF) kernel
   Compute the Reproducing Kernel Hilbert Space (RKHS) distance
6. Select the RBF kernel that maximizes the RKHS distance as the optimal kernel.
7. Given a series of mixture ratios λ, compute its weighted distribution vector $u\lambda$
8. Find the nearest valid probability distribution to vector $u\lambda$, and compute the distance $\hat{\lambda}$ between them.
9. Use the Kernel mean-based gradient thresholder algorithm to compute π, as introduced by [33].

Model training process: The model is trained as in the VFL process: Both the railway and port train their own bottom models to learn the hidden representations of their respective local data. Acting as the active party, the railway plays the dominant role in the framework and consequently maintains the server. This server runs a top model that aggregates the hidden representations received from each participant, i.e., the railway itself and port, and subsequently computes the final output $g(x)$.

Since the training process is a Positive-Unlabeled (PU) learning problem, the learning objective is to minimize the subsequent risk:

$$\mathcal{R}_{pu}(g) = \pi\mathbb{E}_{P(x|Y=+1)}[\ell(g(x),+1)] + \left(\mathbb{E}_{P(x)}[\ell(g(x),-1)] - \pi\mathbb{E}_{P(x|Y=+1)}[\ell(g(x),-1)]\right), \quad (1)$$

where $x \in \mathbb{R}^d$ is the input of the model, and is essentially the concatenation of the hidden representations received from each individual participant. Here, $Y \in \{-1,+1\}$ is the class label, representing the unlabeled and positive samples respectively, $P = \{x_i\}_{i=1}^{n_p}$ is data of size n_p sampled from $P(x \mid Y=+1)$ which equals π, and $U = \{x_i\}_{i=1}^{n_u}$ is data of size n_u sampled from $P(x)$. Also, $\ell(\cdot,\cdot)$ is any trainable surrogate loss function of zero-one loss [34].

In the research context of identifying potential containers, a data imbalance issue exists due to the relatively low proportion of sea-rail intermodal container data in our dataset. Moreover, the selection of containers for railway transportation within our dataset is not entirely arbitrary; it is potentially influenced by specific features like price discounts and other costs. To address this issue, we propose ImbalancednnPUSB. The formulation and derivation of ImbalancednnPUSB is presented as follows:

By oversampling the data in $P = \{x_i\}_{i=1}^{n_p}$ according to $P(x \mid Y=+1)$, we derive $P_o = \{x_i\}_{i=1}^{m_p}$ where $m_p \ll n_p$ and π' is a new class prior around 0.5. The method leaves the class conditional probability unchanged [27], so $P(x \mid Y=+1, o=+1) = P_{\text{balanced}}(x \mid Y=+1, o=+1)$ and $P(x \mid Y=-1) = P_{\text{balanced}}(x \mid Y=-1)$, where $o \in \{+1,0\}$ denotes whether x is labeled (+1) or not (0).

The pseudo-classification risk can be defined as:

$$\mathcal{R}(g) = \pi'\mathbb{E}_{P(x|Y=+1,o=+1)}[\ell(g(x),+1)] + \frac{1-\pi'}{1-\pi}\left[\mathbb{E}_{P(x)}[\ell(g(x),-1)] - \pi\mathbb{E}_{P(x|Y=+1,o=+1)}[\ell(g(x),-1)]\right] \quad (2)$$

Next, we need to optimize an empirical estimation of risk $\hat{\mathcal{R}}(P,U)$, which is

$$\hat{\mathcal{R}}(P,U) = \pi'\hat{\mathbb{E}}_{P(x|Y=+1,o=+1)}[\ell(g(x),+1)] + \frac{1-\pi'}{1-\pi}\left[\hat{\mathbb{E}}_{P(x)}[\ell(g(x),-1)] - \pi\hat{\mathbb{E}}_{P(x|Y=+1,o=+1)}[\ell(g(x),-1)]\right]_+ \quad (3)$$

$$\hat{\mathcal{R}}(P,U) = \frac{\pi'}{n_p}\sum_{x_i\in P}\ell\left(g\left(x_i\right),+1\right) + \max(0,\hat{\mathcal{R}}_N(P,U)) \quad (4)$$

$$\hat{\mathcal{R}}_N(P,U) = \frac{1-\pi'}{n_u(1-\pi)}\sum_{x_i\in U}\ell(g(x),-1) - \frac{(1-\pi')\pi}{n_p(1-\pi)}\sum_{x_i\in P}\ell\left(g\left(x_i\right),-1\right) \quad (5)$$

$$\hat{g} = \arg\min_{g\in\mathcal{H}}\left[\hat{\mathcal{R}}(P,U) + R(g)\right] \quad (6)$$

In these formulas, $\ell(\cdot,\cdot)$ is $\ell(g(\boldsymbol{x}),+1)) = -\log(g(\boldsymbol{x}))$ and $\ell(g(\boldsymbol{x}),-1) = -\log(1-g(\boldsymbol{x}))$ .The risk minimization problem can be defined as Eq. (6).

R denotes a regularization term, and $\mathcal{H}$ signifies the hypothesis class, specifically referring to DNNs in this instance. The procedures of ImbalancednnPUSB are outlined in Algorithm 2.

Algorithm 2 ImbalancednnPUSB

Input: Training data P and U
Parameter: max_epoch, class priors π and π'
Output: $\hat{g}(x;\theta)$

```
Let A be an SGD-like optimizer and t = 0.
while t < max_epoch do
    Shuffle P and U into b mini-batches, each represented as P_i and U_i respectively;
    for i = 1 to b do
        if R̂_N(P_i, U_i) ≥ 0 then
            Set gradient ∇_θ R̂(P_i, U_i);
            Update θ by A
        else
            Set gradient −∇_θ R̂_N(P_i, U_i)
            Update θ by A
        end if
    end for
end while
```

The estimator of the density ratio r is $\hat{r} = \frac{1}{\pi}\hat{g}$, and the choice θ_π amounts to classifying the top - π test data as positive after ranking the inputs by $\hat{r}$. Then, we can derive the final classifier $h(x) = \text{sign}\left(r(x) - \theta_\pi\right)$.

Model evaluation and explanation: In these steps, once the evaluation metric reaches the predefined threshold value, the Shapley Additive Explanations (SHAP) method [35] is employed to elucidate feature importance. From a business standpoint, we analyse the identified potential containers without discounts in terms of the disparities in rail and road transport indicators.

4 FRAMEWORK EVALUATION USING PRACTICAL DATA

4.1 Data Description

The data in this study were drawn from multiple sources. An extensive overview of multiple datasets, including their description, record count, and ownership, is

presented in Tab. 1. We integrated all the data from the three sources, as shown in Fig. 1. The container-related data, including basic information, shipping schedules, stack storage, and operational data, were all collected from the container management system of the port. The railway has two transport stations for containers in this port. The data collection covered the period from June 2022 to July 2023.

The container truck road transportation data was derived from Baidu Maps' truck route planning service, using the primary truck models for container transportation and their respective destinations as parameters.

The container railway transportation data, including daily demands, waybills, and trajectory data, was collected from the China Railway Research Institute's rail freight ticketing system. Notably, this data aligns with the containers transferred by railway as extracted from the port container system.

Table 1 Data sources

Dataset	Fields	Number of Records	Data Owner
Container Basic Information Dataset	Container ID, Type, Size, Weight, Goods Description, Trade Type, etc.	7322158	Port System
Shipping Schedules Dataset	Estimated & Confirmed Arrival Times, Work Start & Completion Times, Departure Time, etc.	1048575	Port System
Stack Storage Dataset	Stack Entry & Departure Times, etc.	9414876	Port System
Container Operation Dataset	Destination, Dispatch Time, Mode of Transportation, etc.	485792	Port System
Container Truck Road Transportation Dataset	Transportation Distance, Fuel cost, Toll Fee, Freight Charges, Duration of Transportation, etc.	150	Baidu Maps
Container Railway Transportation Dataset	Departure & Arrival Stations, Distance Covered, Freight Charges, Discount Policy, Transportation Duration, etc.	76840	China Railway Research Institute

4.2 Data Preprocessing

The data preprocessing in this study includes the following four steps: data cleaning, data integration, data filtering, and feature augmentation. Please note that we conduct the same data preprocessing for datasets from the port and railway sectors, but these are independent tasks performed separately.

In data cleaning, the data is deduplicated, and missing values are filled in. In data integration, container basic information, shipping schedules, stack storage details, and operation data are interconnected using fields like container numbers and bill of lading numbers. Additionally, date fields are converted into features, and categorical fields are presented using one-hot encoding. For the railway sector, the Container Railway Transportation Dataset was already cleaned and integrated upon acquisition.

In data filtering, we chose 150 railway stations as possible destinations for the potential containers. These are the top 150 destination stations for containers, identified from the port's practical data for 2022 and 2023. Containers transported from the port by rail were chosen as positive samples based on these destinations, and their corresponding transportation data were identified in the Container Railway Transportation Dataset. Containers that were transported from the port by modes other than rail were filtered based on their destinations to determine possible railway stations for each container's arrival. Then, considering loading and unloading restrictions, potential containers for transportation were selected. The results of this selection serve as unlabeled samples. In feature augmentation, all samples were supplemented with road transportation features from the Container Truck Road Transportation Dataset. For the unlabeled samples, railway transportation features were supplemented based on the data of the railway arrival and departure stations inferred from their destinations according to the Container Railway Transportation Dataset.

After the abovementioned procedures, the data from multiple sources were integrated into a single dataset. According to the source of data features, the dataset can be vertically split into two portions: data features owned by the port and data features owned by the railway, as shown in Tab. 2.

Table 2 Data features and examples

Feature	Values	Data Owner
Cargo weight	25.5 t, 26.3 t, …	Port
Arrival interval	8.95 h, 6.61 h, …	Port
Wait interval	3.13 h, 1.38h, …	Port
Work interval	14.14 h, 6.70 h, …	Port
Leave interval	2.80 h, 1.73 h, …	Port
Transport interval	4.37 min, 8.20 min, …	Port
Stack interval	249.44 h, 98.29h, …	Port
Container type	HC, RH, FR, RF, RH, TK, …	Port
Container size	20 ft, 40 ft, …	Port
Road transportation distance	580.26 km, 149.08 km, …	Port
Road transportation time	7.16 h, 1.93 h, …	Port
Road fuel cost	303.58 CNY, 77.99 CNY, …	Port
Road tolls	1047 CNY, 215 CNY, …	Port
Road total cost	3261.04 CNY, 763.26 CNY, …	Port
Empty container	E, F, …	Port
Trade type	D, F, …	Port
Rail transportation distance	825 km, 174 km, …	Rail
Rail transportation time	10.31 h, 2.18 h, …	Rail
Rail total cost	3067.6 CNY, 994.2 CNY, …	Rail
95306 rail freight cost	3744.5 CNY, 853 CNY, …	Rail
Discount	1439 CNY, 430.5 CNY, …	Rail

Considering computational constraints and the large size of the whole dataset, our study selectively employed a subset of the available data. We adopted a stratified sampling approach focusing on data from June 2022 to April 2023 to construct the training dataset. This dataset comprised 74,658 records, of which 5,692 were positive samples. We extracted partial data from May to July 2023 to construct the test dataset with 5,258 records, of which 2,127 were positive samples.

4.3 Experimental Setup

In our proposed framework, the classification model can be integrated with any other PU learning method utilizing neural networks. Therefore, we compare the

performance of our proposed method, Imbalancednn PUSB, against other established models such as nnPU, nnPUSB, and ImbalancednnPU. In imbalanced methods, we use SMOTE [36] for oversampling.

We firstly compare the performance of each method between the central and federated experimental settings. Secondly, we compare performance among different methods in the federated setting. For fair comparison, all models adopt consistent network architecture and use the Adam optimization technique. The comprehensive network comprises five hidden layers, each containing 32 neurons. In the federated framework, the top model is structured with three layers, while guest models incorporate two layers. Hyperparameters, specifically the learning rate and weight decay, were tuned from a grid of $\{0.01, 0.005, 0.002, 0.001\}$. We set the number of epochs to be 100. All other hyperparameters within the network remain at their default settings.

Precision, Recall, and F1 Score are the most commonly used metrics to evaluate the model performance and reliability. This study chooses these three metrics to critically compare and assess the identification results of various models, yielding a comprehensive understanding of the model's performance.

4.4 Results and Discussion

Regarding the issues of data imbalance and selection bias, we compare the performances of four models: nnPU, nnPUSB, ImbalancednnPU, and ImbalancednnPUSB. nnPU and nnPUSB do not consider the data imbalance issue, while others do. nnPU and ImbalancednnPU do not consider the selection bias issue, while others do.

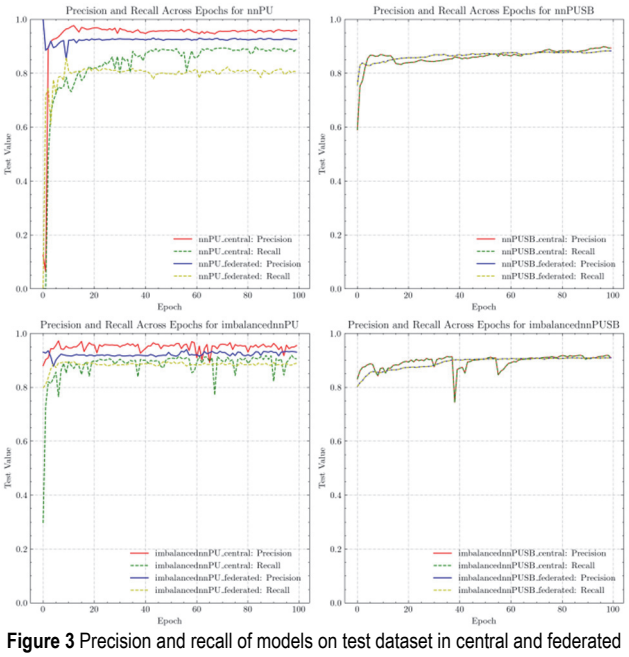

Figure 3 Precision and recall of models on test dataset in central and federated settings

We first evaluate the four models in the central and federated experimental settings. The processes of these models are similar, resulting in comparable resource consumption. Fig. 3 demonstrates the performance of the four models in terms of Precision and Recall in the central and federated settings. For nnPU and ImbalancednnPU, the results in the central setting markedly surpass those in the federated setting. This could be attributed to the reduced parameter count in the federated setting, given the identical network architecture. Similarly, nnPUSB and Imbalanced nnPUSB demonstrate marginally better performance in the central experimental setting than in the federated setting. We observe a similar pattern in the F1 Score, as presented in Fig. 4.

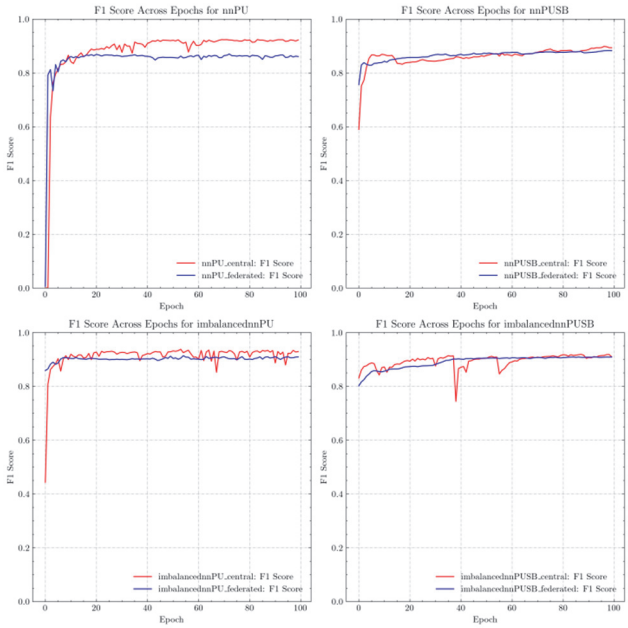

Figure 4 F1 Score of models on test dataset in central and federated settings

In the federated setting, the models considering the data imbalance issue usually demonstrated better performance than their standard counterparts. Specifically, nnPU and ImbalancednnPU achieved almost the same precision, but ImbalancednnPU outperformed nnPU in terms of recall. Similarly, ImbalancednnPUSB consistently outperformed nnPUSB in terms of both precision and recall. Please note that ImbalancednnPUSB exhibits an overlap of its precision and recall curves, indicating a harmonious balance in model performance.

Compared to ImbalancednnPU, ImbalancednnPUSB exhibited a slight dip in precision but compensated with a higher recall. Given that the potential data selection bias (the selection of positive data samples, data associated with transportation via railways, might be biased by price discount), recall would be more valuable than precision in this specific context. Therefore, ImbalancednnPUSB's performance is better than ImbalancednnPU.

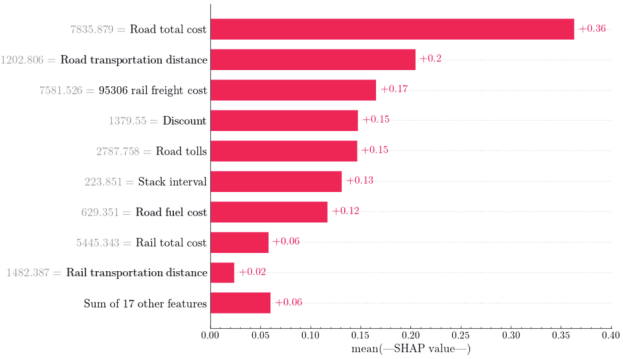

Figure 5 Feature importance based on mean shap value

Fig. 5 presents ImbalancednnPUSB's features ordered by importance using SHAP, suggesting that container identification is significantly influenced by road total cost, road transportation distance, 95306 rail freight cost, and discount policies, among others.

To lower the influence of price discounts on identifying containers and to better understand the underlying business rationales, we randomly selected a batch of unlabeled data as the input into the model, with a focus on those containers identified as having potential value but no price discount. This is because containers without price discounts hold higher value for the railway than containers the railway tries to attract through a price discount policy.

Fig. 6 is a scatter plot showing the difference in distance and cost between rail and road transportation, colored by the time difference between them. Practitioners typically believe that time difference is more important than distance difference to identify potential containers for railways to transport. In this test dataset, the time difference is between 40 hours and 100 hours for potential containers, indicating that railway transportation providers should try to attract containers falling in this time difference range.

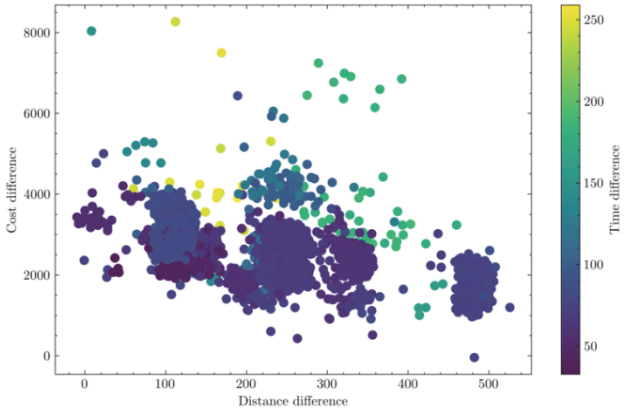

Figure 6 Features distribution of potential containers without price discounts

Containers with too high a time difference, i.e., higher than 100 hours, are unlikely to be chosen by the port for railway transportation. Containers with too small a time difference, i.e., lower than 40 hours, are unlikely to be chosen for railway transportation either. This phenomenon is similar to what is observed in ecommerce, where consumers would pay more attention to delivery time than the distance between sellers and buyers.

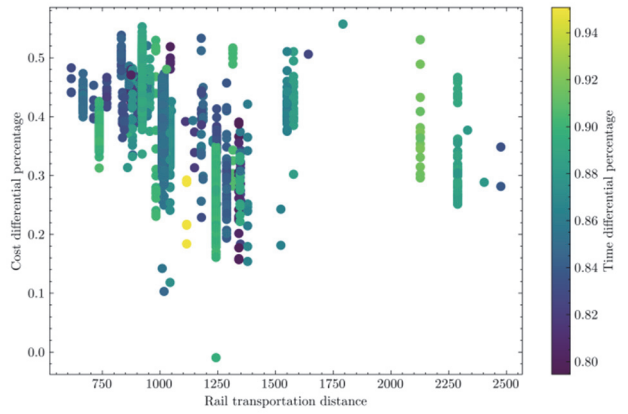

Figure 7 Features of potential containers without price discount based on arrival station

In Fig. 7, the horizontal axis represents the rail transportation distance, with each distance corresponding to the same arrival station. The vertical axis shows the total cost difference compared to road transportation. We can see that the cost difference is higher, typically from 30% to 50%, for closer destinations than for farther destinations, where the cost difference ranges from 15% to 40%. For short-distance railway stations between 500 km and 1000 km away from the port, the average cost difference is 44.75%, and this number is 36.75% for long-distance railway stations between 1000 km and 1750 km away.

This finding implies that the competition is more intense between railway and road when the delivery route is longer; thus, the total cost difference is higher for railway transportation providers if the containers' destination is closer. For destinations between 1000 km and 1750 km, the cost discount is relatively small for railways. This finding has implications for railways and can help to make pricing policies for railway transportation providers to attract potential containers.

From the time perspective, we analyse the time differential percentage for short-distance stations and long-distance stations, and find they are similar: 88.12% and 85.94%, respectively. This indicates that there exists no significant difference in time differential percentage between short-distance stations and long-distance stations.

5 CONCLUSION

Multimodal transportation has been a promising direction in both practice and academia. Original data plays a key role in this context. However, the implementation of data protection laws has challenged traditional techniques for data analysis among multiple parties. The main goal of this study was to design the privacy-preserving framework VFed-PU for potential container identification. This study is the first to combine PU learning and VFL. To address the issues of data imbalances and selection biases in this context, we propose a novel method, ImbalancednnPUSB, to be used during model training. We used real-world data from ports and railways to test the performance of the proposed model. The experimental results demonstrate that our framework can effectively identify the containers that have arrived at the port and are awaiting delivery and could effectively be transported by railway.

The model performance in the federated setting is comparable to that in the central experimental setting, but in the former, parties can simultaneously preserve their data privacy since data are only trained inside the party that owns the data. In the federated model setting, only parameters or weights, instead of original data, are transferred among different parties, thus satisfying data privacy-preserving requirements. The proposed framework can achieve precision and recall of about 90%, indicating an effective balance of data analysis, model training, and data privacy preservation.

The contribution of this study is two-fold. First, we contribute to the literature on freight demand forecasting by mining micro-level data to identify potential containers in the context of sea-rail intermodal transportation. We also quantify the feature importance, thereby generating a detailed and actionable understanding of freight demand in this specific domain. This micro-level approach transcends traditional macroeconomic analyses, offering a data-driven

method to estimate the specific needs of container transportation. Second, the proposed framework offers a novel approach for railways to identify containers with high value, but more important, preserve data privacy by training the model with data inside the data-owning party; only parameters (rather than raw data) are transferred among different parties.

In light of the characteristics identified in potential containers without price discounts, railways can tailor their pricing strategies and speed up services to attract ports to ship these containers by rail. This can facilitate strategic decision-making for railway resource planning, thereby optimizing operational efficiency. Implementing this framework is expected to significantly bolster the volume of sea-rail intermodal transportation, contributing to the advancement of integrated freight networks. This, in turn, could lead to enhanced sustainability in transport logistics by optimizing resource utilization and minimizing wasteful practices.

This study is not without limitations. First, not all of the collected data was used because of limitations on computational resources. Despite this, the experimental results still demonstrate the good performance of our framework. Second, it would be fruitful to explore federated settings with multiple active parties (e.g., several ports and railways) or data with greater heterogeneity. We leave this to future research.

Acknowledgements

Funding: This work was supported by the National Natural Science Foundation of China (Grant No. 52172311 and Grant No. U2268202).

6 REFERENCES

[1] Shi, J., Jiao, Y., Chen, J., & Zhou, S. (2023). Construction of resilience mechanisms in response to container shipping market volatility during the pandemic period: From the perspective of market supervision. *Ocean & Coastal Management*, *240*, 106642.

[2] Archetti, C., Peirano, L., & Speranza, M. G. (2022). Optimization in multimodal freight transportation problems: A Survey. *European Journal of Operational Research*, *299*(1), 1-20.

[3] Mordor Intelligence Research & Advisory. (2023). *Intermodal Freight Transportation Market Size & Share Analysis - Growth Trends & Forecasts (Intermodal Freight Transportation Market).* Mordor Intelligence. https://www.mordorintelligence.com/industry-reports/intermodal-freight-transportation-market

[4] Al Hajj Hassan, L., Mahmassani, H. S., & Chen, Y. (2020). Reinforcement learning framework for freight demand forecasting to support operational planning decisions. *Transportation Research Part E: Logistics and Transportation Review*, *137*, 101926. https://doi.org/10.1016/j.tre.2020.101926

[5] Patil, G. R. & Sahu, P. K. (2016). Estimation of freight demand at Mumbai Port using regression and time series models. *KSCE Journal of Civil Engineering*, *20*(5), 2022-2032.

[6] Tsolaki, K., Vafeiadis, T., Nizamis, A., Ioannidis, D., & Tzovaras, D. (2022). Utilizing machine learning on freight transportation and logistics applications: A review. *ICT Express*.

[7] Li, S., Lv, H., Wang, Y., & Ni, S. (2021). Empty Car Distribution Considering Timeliness Requirement at Chinese Railways. *Journal of Transportation Engineering Part A-Systems*, *147*(10), 04021069. https://doi.org/10.1061/JTEPBS.0000547

[8] Yang, Y. & Yu, C. (2015). Prediction models based on multivariate statistical methods and their applications for predicting railway freight volume. *Neurocomputing*, *158*, 210-215. https://doi.org/10.1016/j.neucom.2015.01.046

[9] Khan, M. Z. & Khan, F. N. (2020). Estimating the demand for rail freight transport in Pakistan: A time series analysis. *Journal of Rail Transport Planning & Management*, *14*, 100176. https://doi.org/10.1016/j.jrtpm.2019.100176

[10] Zhao, L., Cao, N., & Yang, H. (2023). Forecasting regional short-term freight volume using QPSO-LSTM algorithm from the perspective of the importance of spatial information. *Mathematical Biosciences and Engineering*, *20*(2), 2609-2627.

[11] Salais-Fierro, T. E., & Martínez, J. A. S. (2022). Demand Forecasting for Freight Transport Applying Machine Learning into the Logistic Distribution. *Mobile Networks and Applications*, *27*(5), 2172-2181. https://doi.org/10.1007/s11036-021-01854-x

[12] Lu, C., Fu, S., Fang, J., Huang, J., & Ye, Y. (2021). Analysis of factors affecting freight demand based on input-output model. *Mathematical Problems in Engineering*, *2021*, 1-19.

[13] Nuzzolo, A. & Comi, A. (2014). Urban freight demand forecasting: a mixed quantity/delivery/vehicle-based model. *Transportation Research Part E: Logistics and Transportation Review*, *65*, 84-98.

[14] Feng, F., Li, W., & Jiang, Q. (2018). Railway freight volume forecast using an ensemble model with optimised deep belief network. *IET Intelligent Transport Systems*, *12*(8), 851-859.

[15] Wang, P., Zhang, X., Han, B., & Lang, M. (2019). Prediction model for railway freight volume with GCA-genetic algorithm-generalized neural network: empirical analysis of China. *Cluster Computing*, *22*, 4239-4248.

[16] Yang, Q., Liu, Y., Chen, T., & Tong, Y. (2019). Federated machine learning: Concept and applications. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *10*(2), 1-19.

[17] Liu, Y., Kang, Y., Zou, T., Pu, Y., He, Y., Ye, X., Ouyang, Y., Zhang, Y.-Q., & Yang, Q. (2022). Vertical federated learning. arXiv Preprint arXiv:2211.12814.

[18] Cheng, K., Fan, T., Jin, Y., Liu, Y., Chen, T., Papadopoulos, D., & Yang, Q. (2021). Secureboost: A lossless federated learning framework. *IEEE Intelligent Systems*, *36*(6), 87-98.

[19] Feng, Z., Xiong, H., Song, C., Yang, S., Zhao, B., Wang, L., Chen, Z., Yang, S., Liu, L., & Huan, J. (2019). Securegbm: Secure multi-party gradient boosting. arXiv Preprint arXiv:1911.11997, 1312-1321. https://doi.org/10.48550/arXiv.1911.11997

[20] Hou, J., Su, M., Fu, A., & Yu, Y. (2021). Verifiable privacy-preserving scheme based on vertical federated random forest. *IEEE Internet of Things Journal*, *9*(22), 22158-22172.

[21] Romanini, D., Hall, A. J., Papadopoulos, P., Titcombe, T., Ismail, A., Cebere, T., Sandmann, R., Roehm, R., & Hoeh, M. A. (2021). Pyvertical: A vertical federated learning framework for multi-headed splitnn. arXiv Preprint arXiv:2104.00489.

[22] Fu, C., Zhang, X., Ji, S., Chen, J., Wu, J., Guo, S., Zhou, J., Liu, A. X., & Wang, T. (2022). Label inference attacks against vertical federated learning. *31st USENIX Security Symposium (USENIX Security 22)*, 1397-1414.

[23] Sundar, A. P., Li, F., Zou, X., & Gao, T. (2024, July). Toward Multimodal Vertical Federated Learning: A Traffic Analysis Case Study. *2024 33rd International Conference on Computer Communications and Networks (ICCCN)*, 1-9.

[24] Hussain, B. & Afzal, M. K. (2024). Optimizing Urban Traffic Incident Prediction with Vertical Federated Learning: A Feature Selection based Approach. *IEEE Transactions on Network Science and Engineering.*

[25] Du Plessis, M. C., Niu, G., & Sugiyama, M. (2014). Analysis of learning from positive and unlabeled data. *Advances in Neural Information Processing Systems*, *27*.
[26] Kiryo, R., Niu, G., Du Plessis, M. C., & Sugiyama, M. (2017). Positive-unlabeled learning with non-negative risk estimator. *Advances in Neural Information Processing Systems*, *30*.
[27] Kato, M., Teshima, T., & Honda, J. (2019). Learning from positive and unlabeled data with a selection bias. *International Conference on Learning Representations*.
[28] Su, G., Chen, W., & Xu, M. (2021). Positive-Unlabeled Learning from Imbalanced Data. *IJCAI*, 2995-3001.
[29] Hsieh, Y.-G., Niu, G., & Sugiyama, M. (2019). Classification from positive, unlabeled and biased negative data. *International Conference on Machine Learning*, 2820-2829.
[30] Chen, X., Chen, W., Chen, T., Yuan, Y., Gong, C., Chen, K., & Wang, Z. (2020). Self-pu: Self boosted and calibrated positive-unlabeled training. *International Conference on Machine Learning*, 1510-1519.
[31] Xu, C., Liu, C., Yang, S., Wang, Y., Zhang, S., Jia, L., & Fu, Y. (2022). Split-PU: Hardness-aware Training Strategy for Positive-Unlabeled Learning. *Proceedings of the 30th ACM International Conference on Multimedia*, 2719-2729.
[32] Angelou, N., Benaissa, A., Cebere, B., Clark, W., Hall, A. J., Hoeh, M. A., Liu, D., Papadopoulos, P., Roehm, R., & Sandmann, R. (2020). Asymmetric private set intersection with applications to contact tracing and private vertical federated machine learning. arXiv Preprint arXiv:2011.09350.
[33] Ramaswamy, H., Scott, C., & Tewari, A. (2016). Mixture proportion estimation via kernel embeddings of distributions. *International Conference on Machine Learning*, 2052-2060.
[34] Du Plessis, M., Niu, G., & Sugiyama, M. (2015). Convex formulation for learning from positive and unlabeled data. *International Conference on Machine Learning*, 1386-1394.
[35] Lundberg, S. M. & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.
[36] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321-357. https://doi.org/10.1613/jair.953

Contact information:

Lei HUANG, Professor
(Corresponding author)
School of Economics and Management,
Beijing Jiaotong University,
10044, Beijing, PR China
E-mail: lhuang@bjtu.edu.cn

Deyou JIANG
School of Economics and Management,
Beijing Jiaotong University,
10044, Beijing, PR China
E-mail: 20113053@bjtu.edu.cn

Xiong ZHANG, Associate Professor
School of Economics and Management,
Beijing Jiaotong University,
10044, Beijing, PR China
E-mail: xiongzhang@bjtu.edu.cn

Ying WANG, Associate Professor
School of Economics and Management,
Beijing Jiaotong University,
10044, Beijing, PR China
E-mail: ywang1@bjtu.edu.cn

Tianyang BAI, Engineer
China Waterborne Transport Research Institute,
100088, Beijing, PR China
E-mail: 1031802181@qq.com