High value passenger identification research based on Federated Learning

DOI:
10.1109/ihmsc49165.2020.00032

Document Version
Final published version

Published in:
Proceedings - 2020 12th International Conference on Intelligent Human-Machine Systems and Cybernetics, IHMSC 2020

High value passenger identification research based on Federated Learning

Sien Chen
1. Alliance Manchester Business School, University of Manchester, Manchester, UK
2. Research Institute of Internet Industry, Tsinghua University, Beijing, China
3. Antai College of Economics & Management, Shanghai Jiao Tong University, Shanghai, China
sien.chen@postgrad.manchester.ac.uk

Dong-Ling Xu
Alliance Manchester Business School University of Manchester
Manchester, UK
Ling.Xu@machester.ac.uk

Wei Jiang
Antai College of Economics &Management
Shanghai Jiao Tong University
Shanghai, China
jiangwei@sjtu.edu.cn

***Abstract*—Nowadays, airlines are facing increasingly fierce market competition while ushering in development opportunities. Many scholars researched on airline passenger value using data mining approaches, but the evaluation index of air passenger value in the existing research is based on internal data sources. It is of great importance to blend the external data from third-party under the premise of safe and legal data privacy disclosure to extend the characteristic dimension of their customers. Therefore, this research proposes a novel model that can blend multi-source big data to enrich airline passengers' feature dimensions under the premise of ensuring passengers' information privacy security, and establish the user profile of passengers for accurately identifying the high-value passengers. It is proved that our proposed novel model has better performance compared with the results of the traditional model that only use one party data in terms of Area Under Curve (AUC) and Kolmogorov-Smirnov (KS) value.**

Keywords-Federated Learning; Logistic Regression; High Value Passenger;

I. INTRODUCTION

In recent years, with the wave of economic globalization sweeping all kinds of industries, airlines are facing increasingly fierce market competition while ushering in development opportunities. Under the pressure of market competition, how to mine and predict potential high-value passengers ahead of competitors and transform them into high-value trends has become the main concern of airlines.

The research on passenger value in the aviation industry is mainly focused on two directions, one is based on the improvement of the traditional RFM model, and the other is based on data mining technology. In the research of the RFM model, some scholars studied the traditional RFM index [1][2] or make appropriate supplements and adjustments to the traditional index model. For instance, Yeh et al. Extended the RFM model to the RFMTC model by adding two parameters, namely: time since first purchase and churn probability [3]; Xu et al. established the TRFMZ model on the basis of RFM model [4]. In the research based on data mining technology, many scholars used data mining technology and big data technology such as applying clustering [5][6], neural network [7], Markov chain model [8], data extraction transformation loading technology (ETL), online analysis technology (OLAP) and multi-dimensional data analysis and modeling technology [9] to model and calculate civil aviation passenger value; studied civil aviation potential high value passenger discovery method based on RBM-BPNN model [10]. A new airline customer lifetime value estimation model is proposed, which integrates the customer's social network and flight information. By comparing the performance with the airline customer regression model, it is proved that this method can improve the accuracy and reliability of the model of flight-related factors [11].

However, the evaluation index of air passenger value in the existing research is single, and most of the applications of big data technology are only based on RFM model or improved models based on it, which can not make full use of big data's advantages. Due to the singleness of airline data, it is often difficult to analyze passengers' travel habits and unable to understand passengers' satisfaction with services. Airlines urgently need to blend the external data from third-party enterprises or institutions under the premise of safe and legal data privacy disclosure to extend the characteristic dimension of their customers. Then big data and artificial intelligence technology can be applied to analyze the data, which makes airlines thoroughly understand their users and better control their operations for achieving accurate customized services.

Therefore, the main purpose of this research is to propose a novel model which can blend multi-source big data to enrich airline passengers' feature dimensions under the premise of ensuring passengers' information privacy security, and establish the user profile of passengers for accurately identifying the high-value passengers.

The remainder of the paper is organized as follows. In Section II, we introduce the methodologies that will be used for modeling. In Section III, a novel classification model based on Federated Learning is proposed for high-value passengers identification. In Section IV, the experimental results are shown to prove that our proposed model is effective. Sections V summarizes and analyzes empirical results and discusses the future work.

II. METHODOLOGY

A. Federated Learning (FL)

Federated Learning (FL) was proposed by Google in 2016 [12][13], which can achieve multi-party modeling on the basis of ensuring data security, so as to improve the effectiveness of the model. FL can be combined with other privacy preserving techniques like secure multi-party

978-1-7281-6517-2/20/$31.00 ©2020 IEEE
DOI 10.1109/IHMSC49165.2020.00032

computation [14] and differential privacy [15][16][17]. Besides, FL is divided into three categories based on the distribution characteristics of the data, namely Horizontal Federated Learning, Vertical Federated Learning, and Federated Transfer Learning [18]. We applied Vertical Federated Learning to train the model in this project since the training samples are aligned, different features are aggregated in an encrypted state to increase the feature dimensions to enhance the effectiveness of the model.

B. Logistic Regression (LR)

Logistic Regression algorithm is a classification method which often used in data mining, especially for binary classification problems. The function of LR is a "Sigmoid function" which can map a real number to an interval of zero and one. It assumes that the data obeys the Bernoulli distribution, and the gradient descent is used to solve the parameters by the method of the maximum likelihood function.

III. EXPERIMENT

A. Brief introduction of the original dataset

The numerical experiments are conducted on the datasets from customers with airlines. Data A: Airline internal passenger data owned by party A, consisting of 10000 data instances and each instance has 40 attributes containing flight information and consumer behaviors without labels; Data B: TravelSky data owned by party B, consisting of 10000 data instances with the same ID as Data A but 33 different attributes, and label Y–high-value tag (0/1).

ID	Booking amount	Flight number	member points	…
1	6341.00	34	1253	…
2	0.00	0	0	…
3	823.12	8	0	…
…	…	…	…	…

Data A

ID	Proportion by the airline	Booking amount of Air China	…	Y
1	0.32	12678.23	…	0
2	0.10	9425.89	…	1
3	0.91	1274.67	…	1
…	…	…	…	…

Data B

Figure 1. Empirical Datasets

B. Modeling approach

For data privacy and security reasons, parties A and B cannot share their customer data directly. Based on a vertical federated learning framework[18], a third-party collaborator C is created to help both sides construct a linear regression federation model without a data breach. The parameter exchange in the whole training process is under the homomorphic encryption mechanism[19] that allows computing over encrypted data without access to the secret key.

T is the number of data instances and each has n features. $y \in \{-1,1\}$ as the label. Data A owns a disjoint subset of data features over a group of common sample IDs with Data B. Assuming $X = (X^A, X^B) \in R^{n \times T}$, only B has the labels y. The training objective J of logistic regression for classification is:

$$min_{\omega \in R^n} \frac{1}{T} \sum_i l(\omega; x_i, y_i) \tag{1}$$

Where $\omega = (\omega_A, \omega_B)$ is model parameters, $\omega_A \in R^{n_A}, \omega_B \in R^{n_B}$. x_i is the i-th instance and y_i is its label. $l(\omega; x_i, y_i) = \log(1 + \exp(y_i \omega^\intercal x_i))$ as the negative log-likelihood loss function. In order to the compute the loss function and its gradient directly with additively homomorphic encryption, the Taylor approximation for the loss function[20][21] is adopted:

$$l(\omega; x_i, y_i) \approx \log 2 - \frac{1}{2} y_i \omega^\intercal x_i + \frac{1}{8} (\omega^\intercal x_i)^2 \tag{2}$$

Let $S \sqsubseteq \{1,...,T\}$ be the index set of the chosen mini-batch data instances. The corresponding loss and gradient are given by $loss = F(\omega) = \frac{1}{|S|} \sum_{i \in S} l(\omega; x_i, y_i)$, $g = \nabla F(\omega) = \frac{1}{|S|} \sum_{i \in S} \nabla l(\omega; x_i, y_i)$. By denoting $u_A = \{u_A[i] = \omega_A^\intercal x_i^A, i \in S\}$ for A (similarly u_B for B) and $d = \{d_i, i \in S\}$, $[[\cdot]]$ is the homomorphic encryption symbol, the encrypted loss and gradient is:

$$\begin{aligned}[[loss]] \approx \frac{1}{|S|} \sum_{i \in S} & [[log2]] - \frac{1}{2} y_i ([[u_A[i]]] + [[u_B[i]]]) \\ & + \frac{1}{8} \left([[u_A^2[i]]] + 2u_B[i][[u_A[i]]] \right. \\ & \left. + [[u_B^2[i]]] \right)\end{aligned} \tag{3}$$

$$\begin{aligned}[[g]] \approx \frac{1}{|S|} \sum_{i \in S} [[d_i]] x_i = (& \underbrace{\sum_{i \in S} [[d_i]] x_i^A}_{[[g^A]]}, \underbrace{\sum_{i \in S} [[d_i]] x_i^B}_{[[g^B]]}) \\ and\ [[d_i]] = \frac{1}{4} & ([[u_A[i]]] + [[u_B[i]]] + [[-\frac{1}{2} y_i]])\end{aligned} \tag{4}$$

The whole training process can be divided into the following six steps:

- Step 1: Encrypted entity alignment. Using the encryption-based user ID alignment to confirm and align the common users whose IDs are the same between party A and party B;
- Step 2: Party A and B initialize ω_A, ω_B respectively, collaborator C creates encryption pairs, send the public key to A and B;
- Step 3: Party A computes $[[u_A]], [[u_A^2]]$ and sends to B; B computes $[[d_i]]$ and sends it to A, also sends $[[d_i]], [[loss]], [[g^B + maks^B]]$ to C;
- Step 4: Party A computes $[[g^A + maks^A]]$ and sends to C;
- Step 5: Collaborator C updates the loss, decrypts, and sends the decrypted gradient $g^A + maks^A$ back to party A; then sends $g^B + maks^B$ back to party B;
- Step 6: Party A and B unmask and update the model

parameters ω_A, ω_B with g^A, g^B, respectively. Then back to Step 2 until $loss < \varepsilon$ or iteration ends.

During entity alignment and modeling, Party A and B can train a combined model with data kept locally, and hide their gradients from C by the encrypted random masks. Both unilateral data and joint data are modeled by logistic regression algorithm, and Hetero-LR is used to train the federated model. Unlike traditional logistic regression modeling, the airline company and TravelSky use their own data to train the model together, use encrypted intermediate results to interact, which enables each party to maintain its own model. When the prediction is needed, both models need to be combined for common prediction. The whole process of model training ensures the safety of data and models. Finally, we randomly choose 80% data instances as the training set and the remaining 20% as the test set. The experimental results are presented in the following section.

IV. RESULTS

In this paper, AUC and Kolmogorov-Smirnov (KS) value are used to compare and evaluate the results. AUC is the area under ROC curve and is the standard for evaluating the pros and cons of the binary classification prediction model. KS value indicates the model's ability to distinguish positive samples from negative samples. The greater the KS value, the better the prediction accuracy of the model. Generally speaking, if KS value is larger than 0.4, the model can be considered to be of good accuracy.

The results of experiments are shown as below:

Table 1. Model Comparison

Model	Method	Data	AUC	KS
Unilateral Model	Logistic Regression	Data B	0.78	0.42
Hetero-LR Model	Hetero-LR	DataA &Data B	0.85	0.55

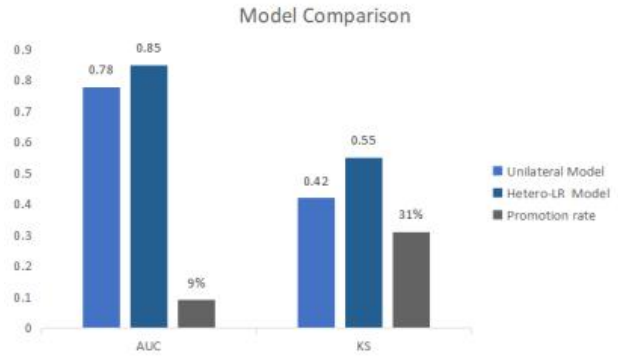

Figure 2. Model Comparison

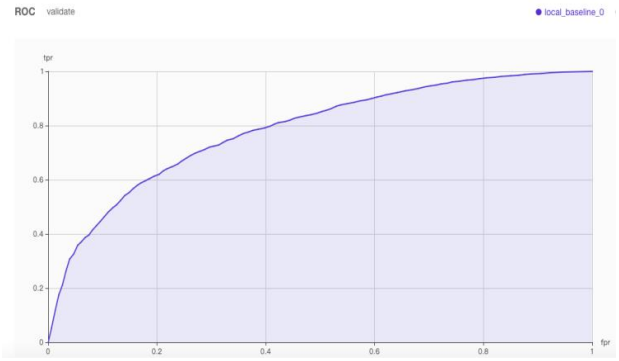

Figure 3. Unilateral Model Results

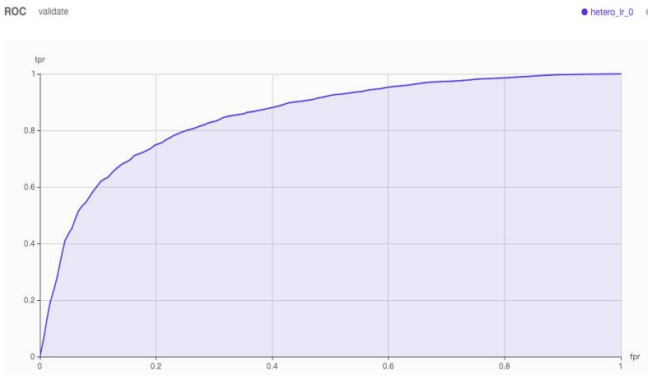

Figure 4. Hetero-LR Model Results

It can be seen from the Table 1 and Figure 3 that AUC of the model with unilateral data and labels is 0.78 and KS value is 0.42. Besides, Table 1 and Figure 4 illustrate the AUC of Hetero-LR model which combines the internal airline passenage data and TravelSky data is 0.85 and KS value is 0.55. It is shown that intuitively in the Figure 2 that the AUC and KS value of Hetero-LR model have improved a lot compared with the results of the model that is only using the characteristics possessed by TravelSky. Specifically, AUC has improved by 9% and KS value has increased by 31%. Judging from these results, the joint model trains more features on the premise of analyzing the same passenger population and ensuring data security, and the model has better effect and more accurate high-value identification of passengers.

V. CONCLUSION

In reality, data island, privacy protection and data security are urgent problems to be solved. In this paper, we proposed a novel classification model based on Federated Learning for high-value passengers identification of the civil aviation industry. To analyse the value of airline passengers, the unilateral model with TravelSky data and joint model combining internal data of airlines and TravelSky data through federated learning are compared. It is concluded that our proposed model not only solves the problem of data island, greatly improves the results of the model, but also better protects user privacy and data security of institutions. The results of the model can be used by airlines to obtain more accurate

identification of high-value passengers. Airlines can have a more comprehensive understanding of the customers, better strengthen the loyalty of the customers.

However, there are some shortcomings in this research. For instance, we only used 73 features to do experiments, which may be extended to more additional features if more parties engage in the modeling process in the future. For the purpose of explanation, this paper has made certain restrictions on the amount of data, so big data will also be applied in modeling.

ACKNOWLEDGMENT

D.L. Xu thanks the support by the European Union's Horizon 2020 Research and Innovation Programme RISE under grant agreement no. 823759 (REMESH).

Wei Jiang thanks the support by National Nature Science Foundation grants 71531010 and 71831006.

REFERENCES

[1] Lumsden, S.A., Beldona, S. and Morrison, A.M., 2008. Customer value in an all-inclusive travel vacation club: An application of the RFM framework. *Journal of Hospitality & Leisure Marketing*, 16(3), pp.270-285.

[2] Yeh, I.C., Yang, K.J. and Ting, T.M., 2009. Knowledge discovery on RFM model using Bernoulli sequence. *Expert Systems with Applications*, 36(3), pp.5866-5871.

[3] Wei, J.T., Lin, S.Y. and Wu, H.H., 2010. A review of the application of RFM model. *African Journal of Business Management*, 4(19), p.4199.

[4] Xu, L., Zhang, S., and Zhang, X. Analysis and Research of airline customer value based on big data. *Information and computer (theoretical version),*2019, 31(23), pp. 109-110.

[5] Dachyar, M., Esperanca, F.M. and Nurcahyo, R., 2019, August. Loyalty Improvement of Indonesian Local Brand Fashion Customer Based on Customer Lifetime Value (CLV) Segmentation. In *IOP Conference Series: Materials Science and Engineering* (Vol. 598, No. 1, p. 012116). IOP Publishing

[6] Seetha, Aditi, and Urjita Thakar. "Finding Customer Loyalty Based on Weighted RFMD Clustering Model." *Available at SSRN 3545085*，2020.

[7] Ayoubi, M., 2016. Customer segmentation based on CLV model and neural network. *International Journal of Computer Science Issues (IJCSI)*, 13(2), pp.31.

[8] Tarokh, M. and EsmaeiliGookeh, M., 2017. A new model to speculate CLV based on Markov chain model. *Journal of Industrial Engineering and Management Studies*, 4(2), pp.85-102.

[9] Dang, Y., Cao, W., and Wang, S., Calculation of civil aviation passenger value based on multidimensional data analysis. *Computer and Digital Engineering*,2017，45(01), pp. 168-171+191.

[10] Xu, T., Liu, Z., and Lu, M., Civil Aviation potential High value passenger Forecast based on RBM-BPNN. *Computer applications and Software,* 2019, 36(09), pp. 58-63.

[11] Çavdar, A. B., and Nilgün, F. "Airline customer lifetime value estimation using data analytics supported by social network information." *Journal of Air Transport Management,* 2018, (67), pp. 19-33.

[12] Konečný, J., McMahan, H.B., Ramage, D. and Richtárik, P., 2016. Federated optimization: Distributed machine learning for on-device intelligence. *arXiv preprint arXiv:1610.02527.*

[13] Konečný, J., McMahan, H.B., Yu, F.X., Richtárik, P., Suresh, A.T. and Bacon, D., 2016. Federated learning: Strategies for improving communication efficiency. *arXiv preprint arXiv:1610.05492.*

[14] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B., Patel, S., Ramage, D., Segal, A. and Seth, K., 2017, October. Practical secure aggregation for privacy-preserving machine learning. *In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (pp. 1175-1191).

[15] McMahan, H.B., Ramage, D., Talwar, K. and Zhang, L., 2017. Learning differentially private recurrent language models. *arXiv preprint arXiv:1710.06963.*

[16] Agarwal, N., Suresh, A.T., Yu, F.X.X., Kumar, S. and McMahan, B., 2018. cpSGD: Communication-efficient and differentially-private distributed SGD. *In Advances in Neural Information Processing Systems* (pp. 7564-7575).

[17] Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I., Talwar, K. and Zhang, L., 2016, October. Deep learning with differential privacy. *In Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308-318).

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

[19] Rivest, R.L., Adleman, L. and Dertouzos, M.L., 1978. On data banks and privacy homomorphisms. *Foundations of secure computation,* 4(11), pp.169-180.

[20] Hardy, S., Henecka, W., Ivey-Law, H., Nock, R., Patrini, G., Smith, G. and Thorne, B., 2017. Private federated learning on vertically partitioned data via entity resolution and additively homomorphic encryption. *arXiv preprint arXiv:1711.10677.*

[21] Yang, K., Fan, T., Chen, T., Shi, Y. and Yang, Q., 2019. A Quasi-Newton Method Based Vertical Federated Learning Framework for Logistic Regression. *arXiv preprint arXiv:1912.00513.*