IEEE Open Journal of the Computer Society

Received 18 October 2023; accepted 18 November 2023. Date of publication 29 November 2023;
date of current version 28 December 2023. The review of this article was arranged by Associate Editor Zhi Liu.

Digital Object Identifier 10.1109/OJCS.2023.3336904

Prediction of Customer Behavior Changing via a Hybrid Approach

NIEN-TING LEE [1], HAU-CHEN LEE [2], JOSEPH HSIN [3], AND SHIH-HAU FANG [1] (Senior Member, IEEE)

[1]Electrical Engineering, Yuan Ze University, Taoyuan 32003, Taiwan
[2]Group Retail Planning Center, Far Eastern, New Taipei 220336, Taiwan
[3]Purdue University, West Lafayette, IN 47907 USA

CORRESPONDING AUTHOR: SHIH-HAU FANG (e-mail: shfang@saturn.yzu.edu.tw).

ABSTRACT This study proposes a hybrid approach to predict customer churn by combining statistic approaches and machine learning models. Unlike traditional methods, where churn is defined by a fixed period of time, the proposed algorithm uses the probability of customer alive derived from the statistical model to dynamically determine the churn line. After observing customer churn through clustering over time, the proposed method segmented customers into four behaviors: new, short-term, high-value, and churn, and selected machine learning models to predict the churned customers. This combination reduces the risk to be misjudged as churn for customers with longer consumption cycles. Two public datasets were used to evaluate the hybrid approach, an online retail of U.K. gift sellers and the largest E-Commerce of Pakistan. Based on the top three learning models, the recall ranged from 0.56 to 0.72 in the former while that ranged from 0.91 to 0.95 in the latter. Results show that the proposed approach enables companies to retain important customers earlier by predicting customer churn. The proposed hybrid method requires less data than existing methods.

INDEX TERMS Customer behavior, customer churn prediction, machine learning, buy till you die (BTYD).

I. INTRODUCTION

In marketing science, customer relationship management is a very important issue for any industry. In a saturated market, the cost of retaining loyal customers is much lower than developing new ones, and stable customers can bring stable revenue to a company [1], [2]. The challenge of predicting customer churn is that every customer has a different reason for churning. As mentioned in [3], there are many reasons for customer churn in a dynamic market environment. A customer may suddenly move out of his or her original place of residence, leave the store where he or she used to spend money, and become a lost customer of the store. It is also possible that a customer has a negative impression of a particular purchase because of a defective product and is unwilling to spend again. This previous work [4], [5] proposes an adaptive learning method, The previous work [6] used a statistical model to observe the customer's consumption cycle and to evaluate the customer's probability of future consumption.

Predicting customer churn can be roughly divided into two categories. The first category is statistical models, which uses customer consumption (visit) behavior characteristics to group customers and assess the probability of customer churn. The second category of research involves [7] using machine learning models to detect individual customer characteristics for predicting customer churn. And the research [8] uses a unique CCP method to deal with the characteristics of whether it is a lost customer. In the first category, [9], [10] used the parameters of the RFM (Recency Frequency Montary) statistical model to divide customers into groups and analyze the characteristics and values of each group of customers in each quarter. On the other hand, Fader et al. [11] used the BTYD (Buy Till You Die) statistical model to calculate the probability of customer alive (P alive). In the second category, researchers have modified or combined traditional machine learning models to predict churn more accurately, such as [3], [12], [13]. Some researchers have applied deep learning algorithms to churn prediction to test the predictive performance of the model in the retail industry, such as [14]. Other researchers have optimized the customer data, divided the customers into groups by the above data as characteristics,

then given the characteristics of each group to analyze the churned customers [2], [15], [16], [17], [18].

This article innovation lies in the integration of statistical model analysis of customer P alive results, followed by customer segmentation based on the observed inter-group mobility. Additionally, we employed a machine learning model to accurately predict the critical churned customers within the dataset, who exhibit high consumption frequency but a gradual decline in P alive values. The proposed algorithm first uses the P alive derived from the BTYD statistical model to dynamically determine the churn line. Compared to the fixed time definition of churned customers, it is less likely to be misjudged as churn for customers with longer consumption cycles. After observing customer churn through clustering over time, the proposed method segmented our customers into four behaviors: new, short-term, high-value, and churn. Finally, machine learning models were used to predict the customers that will be churn, i.e., the customers that are judged by the BTYD model to have low P alive.

Two public datasets were used to evaluate the feasibility of our hybrid approach, an online retail of U.K. gift sellers [19] and the largest E-Commerce of Pakistan [20]. Based on the two datasets having different consumption patterns, we compare the prediction results of 15 machine learning models after cross-validation. In the Online Retail II dataset, the average ROC (Receiver operator characteristic) and AUC (Area Under the Curve) score of the top three models ranged from 0.59 to 0.63, while in the Pakistan's Largest E-Commerce dataset, they ranged from 0.56 to 0.61. In Online Retail II test dataset, the recall of the top three models ranged from 0.56 to 0.72 while that ranged from 0.91 to 0.95 in Pakistan's test data set. Results show that with our suite of hybrid methods for segmenting and observing customer behavior, and the selected machine learning models, the proposed approach enables companies to retain important customers earlier by predicting customer churn.

II. LITERATURE REVIEW

In the literature review, we will start from the origin of statistical models to observe customer behavior. Then, we will pivot to explore machine learning applications in the field of predicting customer behavior in recent years.

The observation of customers' repetitive behavior was first started by the RFM model [21], in which R (Recency) represents the most recent purchase, F (Frequency) represents the number of repetitive purchases, and M (Monetary) represents the average price of the purchase. By using these three elements as coordinates, the RFM model can roughly classify customers into eight groups and differentiate their consumption behavior. In the research of [9], they combined the RFM with the clustering method to investigate the optimal number of subgroups and to analyze the characteristics and value of each group. In recent years, there are still scholars who have made further research on RFM model. For example, [10] also used RFM combined with K-Means algorithm to subdivide the customers and calculated Silhouette Coefficient to verify the effect of clustering. The works in [22] proposed the concept of Customer Lifetime Value (CLV), which explained the value that customers bring to the company during active consumption.

Then the "Buy Till You Die" (BTYD) statistical model emerged, combining the RFM model with CLV estimation. These statistical models can be divided into two types: models that calculate RF (Recency, Frequency) parameters, and models that calculate FM (Frequency, Monetary) parameters. For the RF type, [23] proposed the Pareto/NBD model and [6] proposed the BG/NBD model, which simplified one of the Pareto/NBD calculations and made the model calculation easier to implement. However, this also led to misjudgments in the calculation of customers who did not purchase repeatedly. Thus, [24] proposed the MBG/NBD model, which corrected the the BG/NBD misjudgment situation. For the FM type, Gamma-Gamma model was proposed in [25], making model assumptions on the average spending amount of customers such that the expected value of the spending amount can be calculated.

In addition to the statistical models mentioned above, the rise of artificial intelligence (AI) has led to various application. User behavior is definitely one of the possibilities. The large amount of consumer data has brought the potential of machine learning models for prediction, and therefore many scholars have devoted themselves to research in predicting customer churn. For example, [18] they propose an approach that uses 6 machine learning models to predict changes in customer behavior from two perspectives: individual-level and segment-based. In the research of [3], they optimized 5 machine learning models to achieve 97% accuracy and 84% F-measure (F1 score) on a publicly available competition dataset. while [12] used 8 learning models to predict the results on a carrier dataset and analyzed the impact of data imbalance on the models. The works in [13] proposed a hybrid method combining decision trees and logistic regression. Although these works have investigated the optimization for machine learning models, no more attention is paid to the definition of churn customers in the dataset. The attrition customers was marked by experts in the original dataset. However, it is important to define how these attrition customers are defined for practical application. The research of [2] proposed a segmentation approach in telecommunication dataset and provided some marketing strategies based on the clustering results. [15] studied the data imbalance issue in 3 telco business datasets and combined a framework for predicting churn and churn customer segmentation. The customer segmentation to the insurance industry was investigated in [16], while that to the telecom industry was studied in [17]. These works showed the effectiveness of the customer segmentation in either telecom or insurance industry. However, each industry has different elements of customer churn. This is difficult to apply such methods widely.

Recently, some researchers have hybridized the statistical and the machine learning category. For example, [26] compared the statistical model and machine learning in churn detection and reported that the statistical model is more disadvantaged in the short-term dataset. Moreover, [27] has demonstrated that adding BTYD statistical model parameters to training features allows machine learning models to predict more accurately. However, the determination of churned customers is cut by time. Whether it is cut by multiple time periods to determine, such as [14], or most of the literature uses a period of time to make the determination, they need to rely on manual, to have some understanding of the length of the customer consumption cycle in this dataset (industry).

III. PROPOSED HYBRID APPROACH

This article introduces a innovative hybrid approach that capitalizes on the complementary strengths of statistical and machine learning models to enhance the precision of prediction tasks. The proposed methodology initiates by combining the BTYD model with the K-Means clustering algorithm, establishing a customer churn threshold. Subsequently, statistical models dynamically ascertain churn boundaries by calculating customer survival probabilities, adapting to evolving behavior patterns over time for a more flexible definition of customer churn. Once critical churn customers are identified, common machine learning models are employed, utilizing features derived from the statistical model calculations. Through cross-validation, we ensure robustness in the application of various machine learning models for customer churn prediction. The comparative evaluation of multiple machine learning models aims to identify and leverage the model with the highest predictive accuracy, making our hybrid approach adaptable and effective across diverse datasets. Fig. 1 shows the block diagram of our proposed method, which is divided into three-step instructions.

"The first step is to parameterize the customers through statistical models. We use the MBG/NBD model [24] to evaluate P alive [11] of each customer, and predict the consumption amount of each customer by the Gamma-Gamma model [25]. The second step is to segment customers into groups. We use the P alive and other customer consumption characteristics, and use K-Means to group customers. All customers are classified into four groups, including New customers, Short-term customers, High-value customers, and Churn customers. The third step is to use machine learning techniques to predict changes in the dynamic movement of customers within the group over time, as shown in Fig. 1. The details of the proposed approach are described in the following subsections."

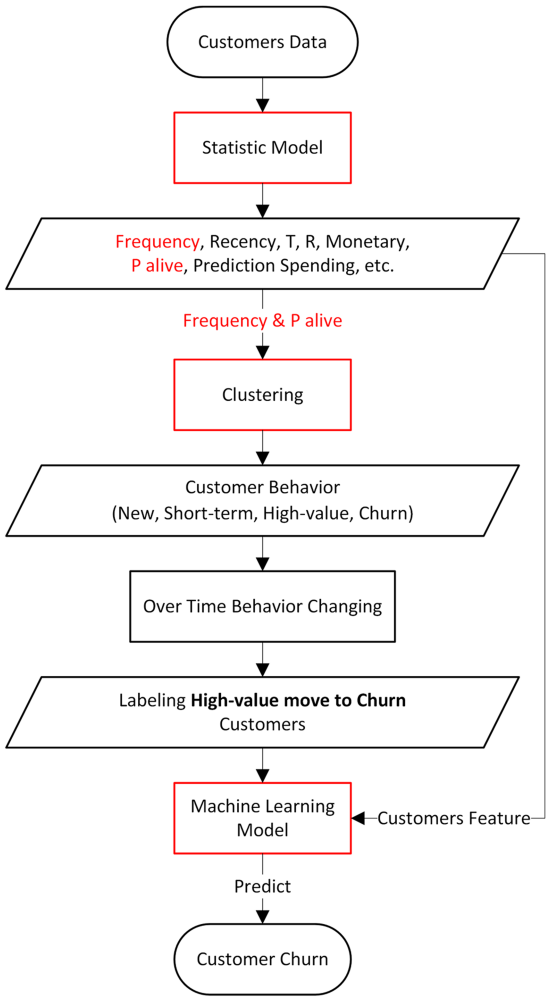

FIGURE 1. Diagram of the proposed hybrid approach.

A. STATISTIC MODELING

Given the customers data, including the customer ID, consumption time, and consumption amount, we can easily convert them into four pieces of information: Frequency (x), consumption interval (t_x), first consumption interval (T), and average consumption amount (Monetary). Then, we will use the MBG/NBD model to evaluate P alive, and use the Gamma-Gamma model to predict the corresponding consumption amount for each customer, respectively.

The input parameters of the MBG/NBD model in BTYD are the first three pieces of information (x, t_x, T), and the output is P alive and the expected number of transactions. The MBG/NBD model incorporates an assumption that there is a decreasing churn rate over time for customers after their first purchase. This solves the problem that customers with no repeat purchases do not churn over time.

The P alive calculation is expressed by (1), where three parameters represent the consumption behavior of the customers, x represents the number of repeated consumptions, t_x represents the number of days between the last consumption and the first consumption, and T represents the number of days between the first consumption and the training deadline of the model. The remaining four parameters of the statistical model, γ, α, a, b, represent the consumption behavior of the group of customers, which be illustrated by the second and

fourth statistical model assumptions in [6], respectively.

$$P(\text{ alive} \mid x, t_x, T, r, \alpha, a, b) = \frac{1}{1 + \frac{\Gamma(a+1)\Gamma(b+x)}{\Gamma(a)\Gamma(b+x+1)} \left(\frac{\alpha+T}{\alpha+t_x}\right)^{r+x}} \tag{1}$$

Eq. (2) was derived to calculate the expected number of transactions in a time period of length t for each customer, in which λ represents the transaction rate of the customers, and p represents the probability of a customer becomes inactive. The above two parameters can be illustrated by the assumptions of the statistical models in [6].

$$E(X(t) \mid \lambda, p) = \frac{1-p}{p}\left(1 - e^{-\lambda p t}\right), \quad \text{where} \lambda \text{is } \gamma, \alpha \text{ and} p \text{ is } a, b \tag{2}$$

The input parameters of the Gamma-Gamma model are Frequency (x) and Monetary, and the output is the prediction of the corresponding consumption amount for each customer. The gamma-gamma model [25] is a model of monetary value, which enables to make inferences about customers' unobserved mean transaction values. In this article, we call this value Prediction Spending. In the x times of consumption, the average consumption amount ($\bar{z}$) of each consumption is represented by z_i. This model assumes that $z_i \sim \text{gamma}(p, \nu)$, where the expected value of z_i can be expressed as $E(Z_i \mid p, \nu) = \zeta = p/\nu$ and $\nu \sim \text{gamma}(q, \gamma)$.

After a series of mathematical derivations in [25], the expected value of customer's consumption amount Z can be calculated by the parameters $p, q, \gamma, \bar{z}, x$, as shown as (3).

$$\begin{aligned} E(Z \mid p, q, \gamma; \bar{z}, x) &= \frac{p(\gamma + x\bar{z})}{px + q - 1} \\ &= \left(\frac{q-1}{px+q-1}\right)\frac{p\gamma}{q-1} \\ &\quad + \left(\frac{px}{px+q-1}\right)\bar{z} \end{aligned} \tag{3}$$

B. CLUSTERING

After P alive and Frequency (x) are calculated by the statistic models, we divide customers into groups through the K-Means grouping algorithm [28]. Based on the above segmentation results, we calculated the percentage of repurchase for each cluster during the other (test) period, called repurchase rate. Such repurchase rates were utilized to define the P alive threshold for customers who had been lost, called P alive churn threshold. Afterwards, the maximum P alive of the clusters with repurchase rate below 50%, is the P alive churn threshold.

The number of clusters of K-Means can be set from an initial value. By gradually increasing the k-value, the above threshold definition will gradually tend to be the same. Therefore, we can set different k-values for different datasets until the P alive churn threshold tends to be the same. By using the P alive churn threshold, customer behavior can be divided into active and inactive customers. Meanwhile, by using frequency threshold, active customers can be divided into new customers and high-value customers, and inactive customers can be divided into short-term customers and churn customers. Over time, customer behavior will change. By increasing or decreasing the number of repurchases, the new P alive would be updated by the statistical model.

Among the four customer groups, the high-value group is the most important one. We define the behavior change of this group of customers as churn customers. Then, we mark such customers in high-value groups as labels for training machine learning models, where the features are outputs of the used statistical models.

C. MACHINE LEARNING

After completing the above customer segmentation and marking the churned customers, we used supervised learning to train machine learning models to predict customer churn. We use all the statistical model parameters in Section III-A with a total of 9 features. The nine features include Frequency(x), t_x, T, Recency(R), Monetary, P alive, Prediction Spending, and expected value of consumption in two period (before and after). Then, we use the mentioned high-value group of customers and the marked churned customers as the training answers, and the 9 features for machine learning models.

There are 15 machine learning models were adopted in the proposed hybrid approach. The 15 machine learning models are not utilized in a blended manner. Instead, we obtain the parameters for each statistical model and carry out individual predictions. The following is a selection of 14 traditional models, as well as the Deep Neural Network (DNN) models that have been commonly used in recent years. These traditional machine learning models for predicting customer churn include commonly employed to predict probabilities and classify customers as either churned or non-churned Logistic Regression (LR) model [29], The predictions from these trees are combined to form the final prediction Random Forest Classifier (RFC) [30], Particularly suitable for handling continuous feature data in classification problemsGaussian Naive Bayes (NB) [31], A gradient boosting framework known for its speed and efficiency, capable of handling large-scale datasets LightGBM (LGB) [32], An ensemble learning technique that typically builds multiple models using bootstrap sampling and combines their predictions through averaging or voting to make the final prediction Bagging Classifier (BGC) [33], Repeatedly divides the dataset into different subsets to make the final prediction Decision Tree Classifier (DTC), An ensemble learning technique often implemented using boosting algorithms, incrementally improving the accuracy of weak learners to build a powerful predictive modelGradient Boosting Classifier (GB), Excelling in large-scale datasets and offering regularization features to prevent overfitting XGBoost (XGB) [34], Analysis achieves classification by projecting features into a lower-dimensional space Linear Discriminant Analysis (LDA) [35], and a linear model that uses L2 regularization to control model complexity and mitigate issues with multicol linearityRidge Classifier

(RC) [36]. In the presentation of our research findings within the manuscript, we account for the heterogeneity across various datasets, recognizing the plausible challenge of data imbalance. Two of the above models, RFC and BGC, are optimized for data imbalance using two different python packages.

RFC were abbreviated as RFCCW1 (balanced weights), and RFCCW2 (balanced subsample weights) using the scikit-learn package [37] while RFC are abbreviated as BRFC (Balanced Random Forest Classifier) using the imblearn package [38]. That is, we have 4 kinds of RFC models. Similar to RFC, BGC was abbreviated as BBGC (Balanced Bagging Classifier) using the balanced weight model of imblearn package and 2 kinds of BGC models were used [39]. For the DNN model, we have also adjusted them for the balanced weights by referring to [40]. Finally, there are 15 machine learning models involved in the model selection of the proposed mechanism

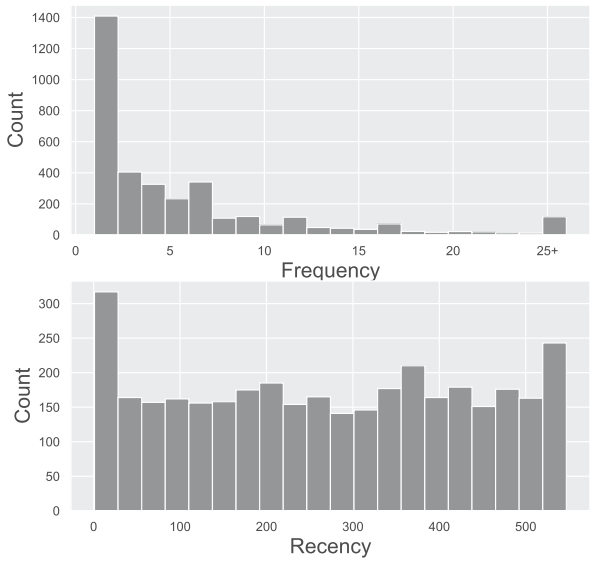

FIGURE 2. Frequency and recency distribution of customers in online retail II.

IV. DATA DESCRIPTION AND EXPERIMENTAL SETUP

A. DATA DESCRIPTION

Obtaining datasets that record customers' long-term consumption behaviors is challenging, and we sought datasets with specific characteristics, such as being from reputable companies, having a sufficient duration of consumer activity, and containing an appropriate number of customers. "We sought two datasets with distinct characteristics, utilizing the diverse customer consumption behavior inherent in both datasets for model analysis. The findings confirm that our research model is capable of adapting effectively to datasets with different characteristics." After an extensive and long-term search and experimentation. This study used two online public databases provided on the Kaggle website to verify the prediction algorithm of this article, namely "Online Retail II" [19] and "Pakistan's Largest E-Commerce" [20].

The Online Retail II dataset is pulled from the UCI Machine Learning Repository. It contains all the transactions occurring for a U.K.-based and registered, non-store online retail between 01/12/2009 and 09/12/2011. The company mainly sells unique all-occasion gift-ware. Many customers of the company are wholesalers. There are 824364 transaction data with 5002 customers in this dataset. Through Fig. 2, it can be observed from the frequency that the number of customers who repeatedly consumed this dataset for more than 25 times is quite significant, and it can be found from the interval recency that almost half of the customers are long-term consumers. This phenomenon confirms the statement that most of the customers are wholesalers as mentioned in the description column.

Pakistan's Largest E-Commerce dataset is the largest retail e-commerce orders dataset from Pakistan. The data was collected from various e-commerce merchants as part of a research study and was first published in the Data Science course at Alnafi (alnafi.com/zusmani). It records from

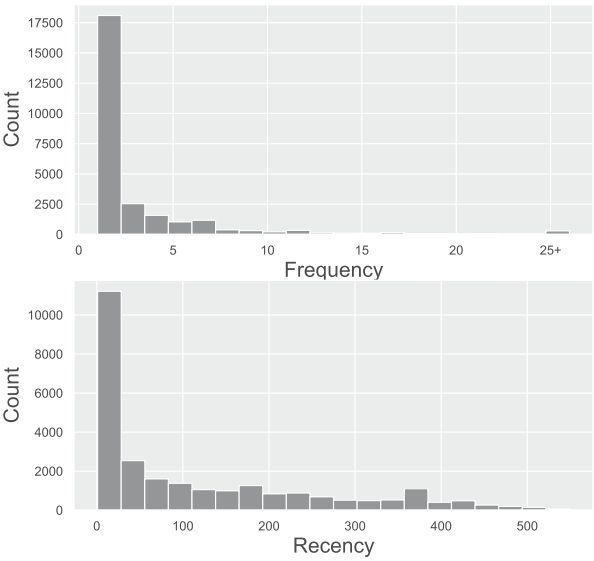

FIGURE 3. Frequency and recency distribution of customers in Pakistan's largest E-commerce dataset.

07/01/2016 to 08/28/2018, a total of 582,281 transaction data with 115081 customers. Part of the data is the order status, including "completed", "cancelled", "order" and "Refund". We only used transaction data with order status "Completed". Filtering out invalid orders by removing the above allows our model to predict purchase behavior without being affected by invalid orders. From the above consumption data and the number of customers, it can be inferred that the number of repeat purchases in this dataset is less than that in Online Retail II. This is confirmed by frequency (the distribution of repeat purchases) in Fig. 3. Also, the major difference with Online Retail II in Fig. 2 is that recency (the length of time

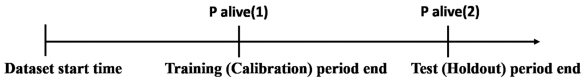

FIGURE 4. Dataset time slice.

between purchases) shows that most customers are focused on short-term spending.

B. EXPERIMENTAL SETUP

Fig. 4 shows our time slice for the dataset, represented by the time axis. From the Dataset start time to Training (Calibration) period end, we can calculate P alive (1) by the MBG/NBD statistical model. From the Dataset start time to Test (Holdout) period end, we can calculate P alive (2). The P alive (1) in the training period is used to segment the customers, and the P alive (2) in the test period is used to observe the change of customers behaviors. For the time slice length setting, we refer to the average daily consumption of customers from the experimental data set. Our slice training period is one and a half years, and the test period is six months. Detailed reasons will be discussed in Section V.

For the evaluation metrics of machine learning models, we use the common evaluation methods, the four metrics extended by the confusion matrix, which are precision, recall, ROC AUC score [41], and F1 (F-measure). After assessing the performance of three predictive models with better performance on the training set through ROC AUC score and Recall, we will incorporate Precision and F1 in conjunction with the previous two metrics for a more comprehensive observation of the model's training set performance. Finally, Precision and Recall will be objectively examined to observe the results on the test set. We also use cross-validation to investigate the effect of data variation on machine learning models. In the following experiments, the training data are cut into 5 parts and the mean and standard deviation are calculated after cross-validation. This is to investigate whether the model over-fits a particular data and under-predicts the other data.

V. RESULTS AND DISCUSSION

A. RESULTS

The heterogeneity distribution of the parameters p and λ in the model can be used to roughly observe the fit of the data and determine the time split of the dataset. By observing the heterogeneity of the model parameters p and λ, the distribution of the churn rate and the average number of days between consumption can be observed respectively. The p in Fig. 5 shows that most of the customers in this dataset have a low churn rate. A larger value of λ means a shorter consumption period. We can observe that most of the customers' interval distribution is concentrated within 0.075, and the statistical density before 0 reaches 8000, which means that most of the customers have an average purchase interval of more than 13 days, and they do buy back.

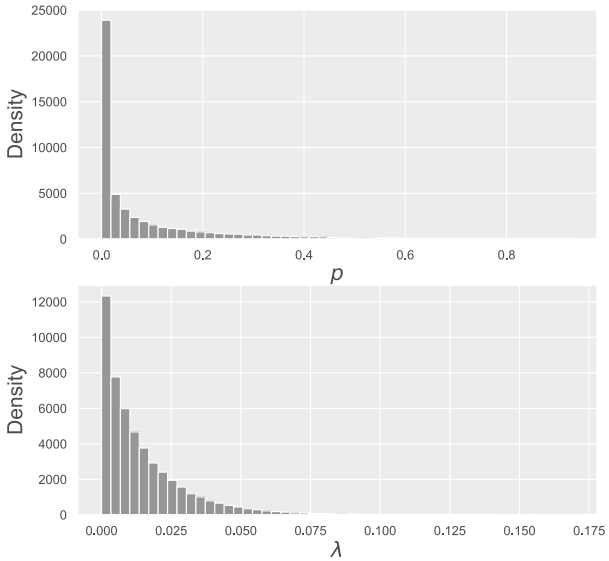

FIGURE 5. Parameters heterogeneity of MBG/NBD model with fitting online retail II dataset.

After knowing the distribution of the model parameters, we calculate the mean of λ to be 0.014, which represents the average of 0.014 visits per day in the dataset. This can be interpreted as about 1.4 visits in 100 days. In order to observe the long-term spending of regular customers during the training period, we need to be able to observe at least 5 times of customer spending. It takes at least 357 days of training to get 5 visits from customers. In order to facilitate the observation, the training period was finally set to one and a half years, or 547 days, which is half a year longer, so that the model can better detect the characteristics of long-term customers. The test period was set to half a year, or 182 days, so that we could confirm that customers would return after observing only one customer's consumption.

Fig. 6 shows the performance of the MBG/NBD model during the training period, where the horizontal axis represents the number of repurchases, the vertical axis represents the number of customers, the blue bar represents the actual number of customers, and the orange bar represents the total number of customers after adding up the repurchase expectations calculated by the model. This figure shows that the actual and predicted numbers are almost similar, which proves that the statistical model successfully detects the consumption pattern of the customers in this dataset.

However, it is difficult to determine whether the model can predict accurately only for the training period data. Therefore, we need to verify whether the model can still predict accurately for the untrained data in the test period. Fig. 7 shows the difference between the predicted and actual values of the statistical model during the test period. In this figure, the horizontal axis represents the number of times the customer spent during the training period, the vertical axis represents the average number of times the customer spent during the

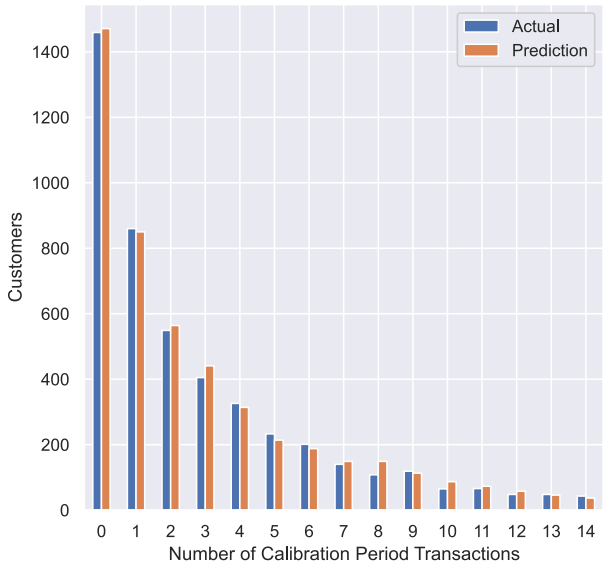

FIGURE 6. MBG/NBD model fitting performance with calibration period transactions in online retail II.

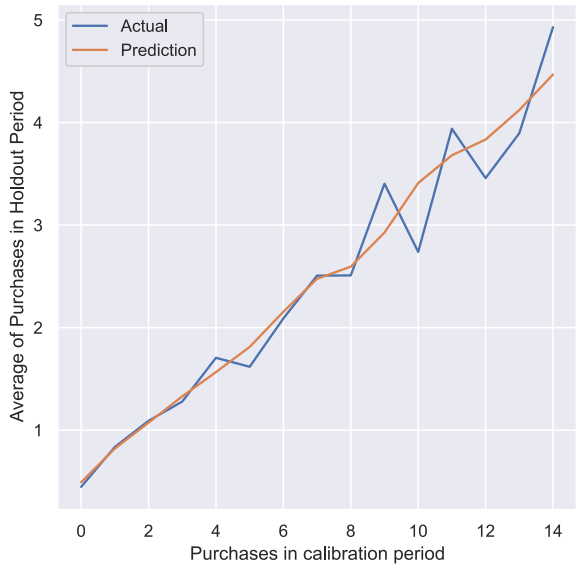

FIGURE 7. MBG/NBD model performance with holdout period purchases in online retail II.

validation period, the blue line is the actual value, and the orange line is the predicted value of the statistical model. Results show that the actual value of the average consumption in the validation period and the predicted value of the average consumption in the training period are almost exactly matched, while the error of the predicted average in the second half of the period is mostly within 1 time. Here we calculate the correlation between the actual and predicted values and get a positive correlation of 0.84. That is, the MBG/NBD model successfully fits this dataset such that we can trust the P alive value to do the clustering.

Next, Fig. 8 shows the clustering results of the entire training customers based on the frequency and P alive values.

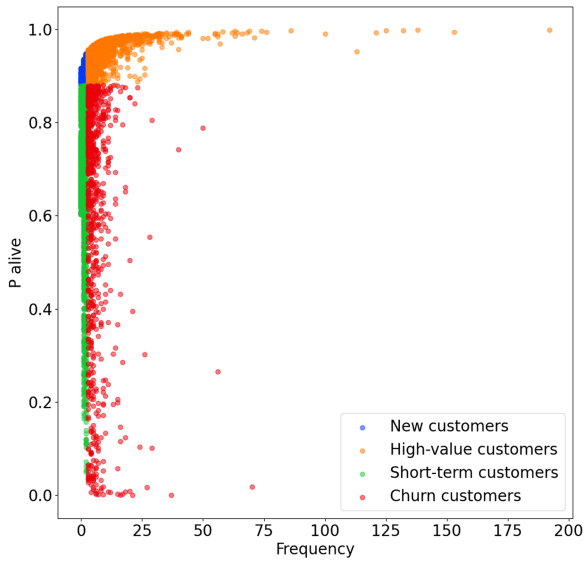

FIGURE 8. Customer segmentation distribution map in online retail II.

There are four clusters: new, high-value, short-term, and churn customers, as denoted by different colors. The majority of customers are concentrated in the upper half of the market, so the churn threshold defined by this study is relatively high, at 0.88. In this retail case, this thesis defines repeat customers as those who have made more than three purchases.

Similarly, through Fig. 8, we can find that the area with higher density of customers is the upper left, which means that most of the customers in this dataset have a stable consumption behavior during the training period. We can also observe from the maximum value of the frequency on the x-axis that the customers in the high-value. The maximum value of the repetition frequency in the x-axis shows that customers in the high-value group spend up to 200 times during the training period. We can then see the red group, which is a churned customer, but there are a few customers with nearly 75 repetitions, so we can judge that this customer is not spending well in the latter part of the training period, and has a high frequency of spending in the early part of the training period (maybe once every other week). The MBG/NBD model judged that the customer had been lost.

Next, Fig. 9 shows the changing behavior of the clustering results after 6 months, especially for the high value customers. In conjunction with the P alive churn threshold defined in this thesis, the statistical model calculates the customer P alive again after the test period. From Fig. 9, we can observe that P alive of those customers who do not have a significant increase in frequency decreases significantly, and therefore moves to the red inverted triangle, the lost customer group. This change in status is the change in customer behavior that is the focus of this article: "customer churn".

Through Fig. 10, we can see the number of people in each group and the proportion of changes. The two groups with the largest number of people and the smaller proportion of changes are short-term and high-value customers. Combined

TABLE 1. Model Training Cross Validation Results in Online Retail II

Model	Precision		Recall		ROC AUC score		F1	
	exp	std	exp	std	exp	std	exp	std
LR	0.278083	0.015622	0.657143	0.053452	0.630817	0.026792	0.379743	0.022807
DNN	0.258944	0.022163	0.542857	0.201131	0.622808	0.032042	0.395610	0.034785
BRFC	0.276999	0.033818	0.585714	0.084649	0.596916	0.018421	0.365889	0.048505

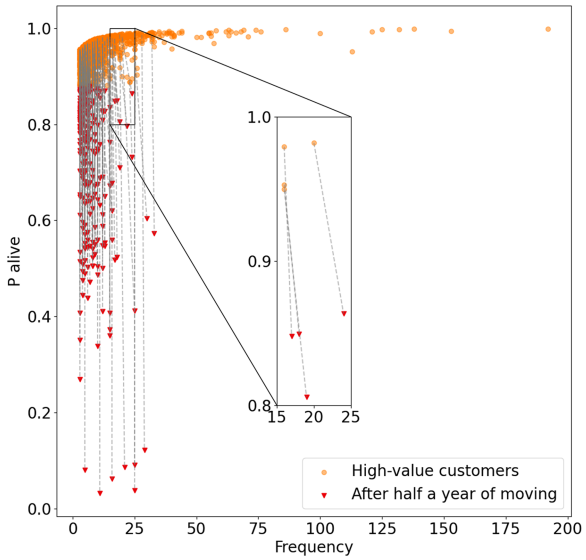

FIGURE 9. Customer segmentation move map in online retail II.

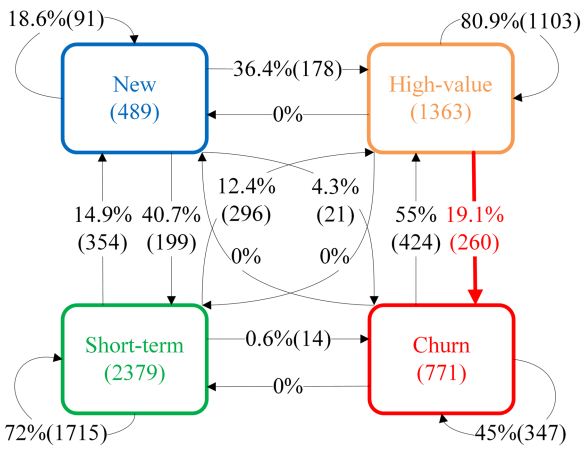

FIGURE 10. Schematic diagram of changes in customer segmentation in online retail II.

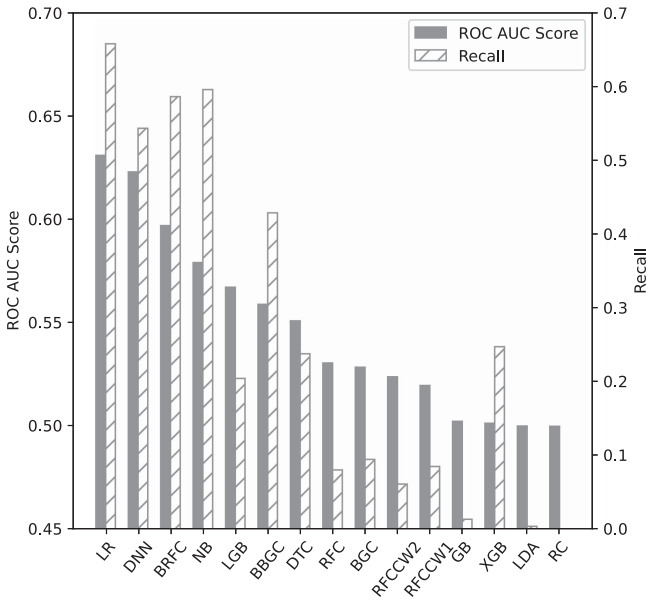

FIGURE 11. Performance of 15 machine learning models trained in online retail II.

TABLE 2. Prediction of Test Data in Online Retail II

Model	Precision	Recall
LR	0.34	0.72
DNN	0.30	0.86
BRFC	0.29	0.64

with Fig. 9, we can find that when high-value customers leave, they will only move to the lost customer group, which accounts for 19.1% of high-value customers. Here we define these customers as churn customers. Fig. 11 shows the prediction results of 15 models for the above-marked churn customers, ranked by the ROC AUC score, with recall assisting to observe the prediction results. From the ranking results, LR, DNN, BRFC, and NB are the top four and their recall is also the top four among all models, which can be said to dominate the prediction results of Online Retail II.

In Table 1, we can see the results of the five cross-validations of the top three models in the above figure, as well as the mean (exp) and standard (std) deviation of each indicator. By observing the precision, we can find that the difference in performance between the three models is not large but the LR is more stable. The standard deviation of the DNN model is larger than the other two models, and its F1 score is the highest among the three models. After the above observations, we found that each of the three models has its own strengths, LR, BRFC is stable, and DNN is more accurate than the first one in some cases. From Table 2, we can see the performance of the three models in the independent test data. Echoing the previous paragraph, DNN has excellent performance in recall in some conditions, even higher than the

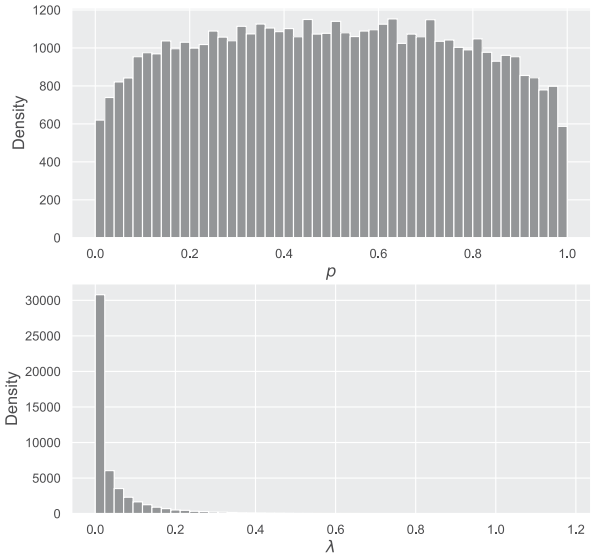

FIGURE 12. Parameters heterogeneity of MBG/NBD model with fitting Pakistan's largest E-commerce dataset.

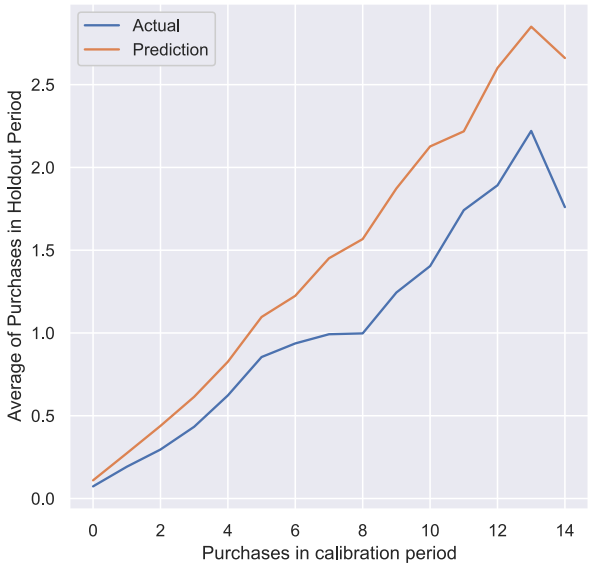

FIGURE 13. MBG/NBD model performance with holdout period purchases in Pakistan's largest E-commerce dataset.

results tested in cross-validation, which confirms the potential of DNN.

B. EVALUATION IN ANOTHER DATASET

To verify the effectiveness of the proposed method, in clustering customers behavior, we evaluated the performance using an alternative online retail dataset on Kaggle, Pakistan's largest E-Commerce dataset. Fig. 12 shows the heterogeneity of the two parameters of the MBG/NBD model when applied to this database. Unlike Fig. 5 in Online Retail II, the distribution of the parameter p is quite even, with a certain density of both low and high drop out rates, which tells us that there are many churned customers in this database. The distribution of λ shows that the density at the 0 position is as high as 30,000, which means that most of the customers will not be able to make their next purchase in the near future. However, the inverse of λ is the average consumption interval of customers, and we can still observe a small number of density bars in the λ distribution around 0.2, so we can know that most of the customers in this dataset have a short repurchase period, about 5 days or more.

Again, we calculate the average of λ as 0.0438, so we need to observe at least 5 consumptions, which takes 125 days. Compared to the previous dataset (about 232 days less), we can infer that most of the customers in this dataset are short-period consumers. However, in order to fairly compare the impact of the two datasets on the statistical model, we also divided the training period into one and a half years (547 days) and the test period of 6 months (182 days).

From Fig. 13, we can verify the difference between the predicted and actual number of customers' consumptions during the test period. However, unlike the previous dataset, it is clear that the statistical model overestimates the customers in this dataset throughout the entire prediction period, and with the red arrow churning percentage in Fig. 15, it can be inferred that the churning of customers in this dataset may be more serious than observed.

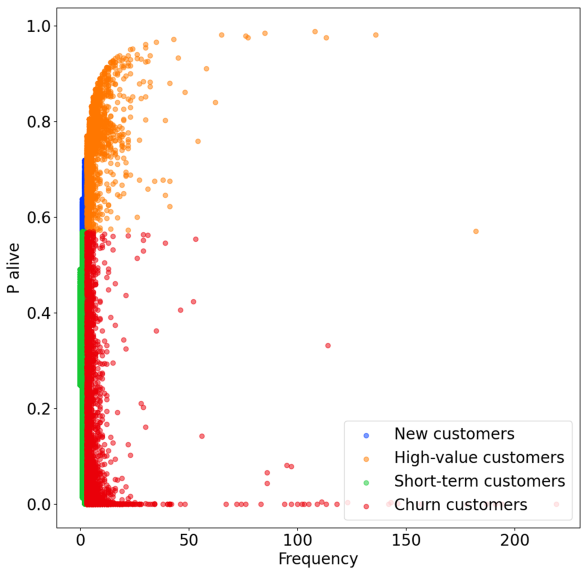

FIGURE 14. Customer segmentation distribution map in Pakistan's largest E-commerce dataset.

After this dataset is divided into groups, the churn boundary is 0.57, and the old and new customer boundaries are the same 3 times, as shown in Fig. 14, which is different from the previous data. Most of the customers are concentrated in the lower half, so the churn boundary determined by the method in this article is also lower. From Fig. 14, it can be roughly deduced that most of the customers in this dataset have already churned, both in the short and long term, and that there are fewer new customers coming

TABLE 3. Model Training Cross Validation Results in Pakistan's Largest E-Commerce Dataset

Model	Precision		Recall		ROC AUC score		F1	
	exp	std	exp	std	exp	std	exp	std
BRFC	0.937229	0.011485	0.580350	0.033078	0.618604	0.039500	0.709404	0.019735
LR	0.946737	0.011891	0.518461	0.042406	0.604091	0.046855	0.685265	0.019360
BBGC	0.929432	0.009718	0.630224	0.010177	0.569786	0.047741	0.766971	0.011466

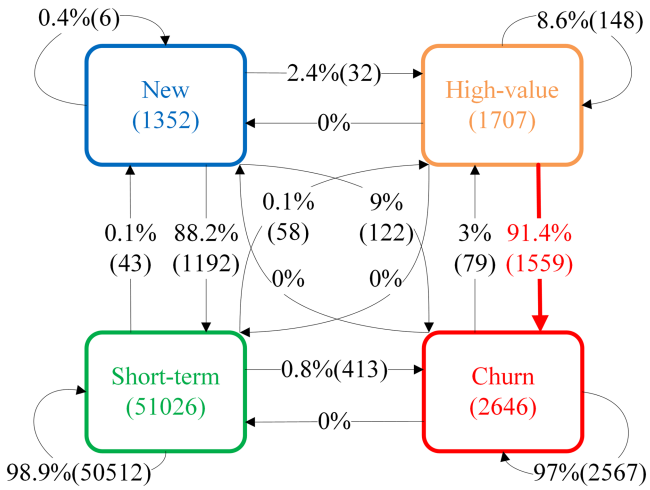

FIGURE 15. Schematic diagram of changes in customer segmentation in Pakistan's largest E-commerce dataset.

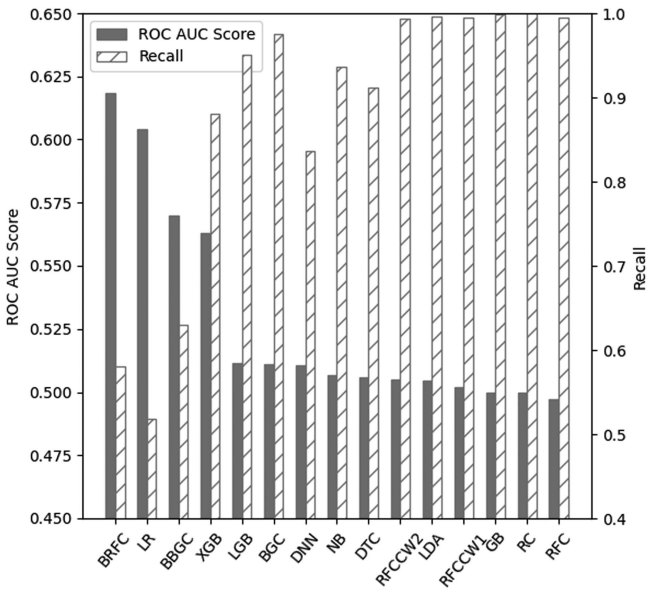

FIGURE 16. Performance of 15 machine learning models trained in Pakistan's largest E-commerce dataset.

to spend money in the near future, so the blue group representing new customers is particularly close to the churn line.

After 6 months of the test period, the red arrows in Fig. 15 show that the high value change is for the churned customer group. As time passes, the P alive of the customers who were in the high value group have more customers drop to below the P alive churn threshold. From this, we can determine that this dataset is losing a large number of customers. Through Fig. 15, it can be observed that in this dataset, a large proportion (91.4%) of customers are lost, which is why the data is unbalanced when the machine learning model captures the characteristics of lost customers. Therefore, as long as the model judges all customer churn, it can have a predicted recall of nearly 91%.

Fig. 16 shows that the top three prediction models ranked by the ROC AUC score have changed. The two models that retain the top three positions are BRFC and LR, and the ROC AUC scores of these two models all remain above 0.6. However, after seeing the performance of recall, the last few models are better. This phenomenon once again reflects the problem of imbalanced training data.

The top three models are shown in Table 3, and we can observe that the standard deviation of the prediction results of these three models is relatively small, which means that the three models do not have a high impact on the data cut. From the precision results, although the models are able to predict more than 90% of the customers correctly, from the recall results, 50% to 60% of the correct answers are really covered, which means that among those who are predicted to be churned by the models, all of them are indeed churned customers. However, there are still nearly 40% of the churned customers not predicted.

TABLE 4. Prediction of Test Data in Pakistan's Largest E-Commerce Dataset

Model	Precision	Recall
BRFC	0.93	0.53
LR	0.95	0.47
BBGC	0.91	0.61

Although the three models in the table are ranked in the top three by the average ROC AUC score, the ultimate goal of predicting churn is to make early countermeasures. Therefore, in this dataset case, a large number of customers are churned and countermeasures are made for all high-value customers.

From Table 4 test data results, we can see that these three models have little impact on the different data, but all belong to the higher precision, lower recall prediction results. In fact, in the data imbalance, the ROC AUC score will be judged as a better result of the forecast. However, as mentioned above, a large number of customers have been observed to be lost during the test period, and it is recommended to make countermeasures for all the remaining high-value customers or to develop new sources of customers.

VI. CONCLUSION

In this article, we propose a hybrid approach using statistic and machine learning models. We mainly use two BTYD statistical models, MBG/NBD and Gamma-Gamma, where MBG/NBD is used to calculate the P alive that defines the churn boundary, and the output parameters of Gamma-Gamma are used to train the machine learning model. With a good fit of the statistical model, it is possible to dynamically define the churn boundary according to different types of consumer behaviors. Two public datasets were used to evaluate the hybrid approach, an online retail of U.K. gift sellers and the largest E-Commerce of Pakistan. The experimental results show that our method defines 19% and 91% of the total number of churned customers in two very different datasets, respectively. Our model achieves a maximum recall of 0.86 and a maximum precision of 0.34 in Online RetailII test data, and a maximum recall of 0.61 and a maximum precision of 0.95 in Pakistan. An alternative advantage of our method is that it only requires the customer's spending history and amount. Considering consumer history as a kind of repetitive behavior, it can be applied to predict any repetitive behavior in diverse industries.

REFERENCES

[1] L. Yan, D. J. Miller, M. C. Mozer, and R. Wolniewicz, "Improving prediction of customer behavior in nonstationary environments," in *Proc. Int. Joint Conf. Neural Netw.*, 2001, pp. 2258–2263.

[2] W. Bi, M. Cai, M. Liu, and G. Li, "A Big Data clustering algorithm for mitigating the risk of customer churn," *IEEE Trans. Ind. Inform.*, vol. 12, no. 3, pp. 1270–1281, Jun. 2016.

[3] T. Vafeiadis, K. I. Diamantaras, G. Sarigiannidis, and K. C. Chatzisavvas, "A comparison of machine learning techniques for customer churn prediction," *Simul. Modelling Pract. Theory*, vol. 55, pp. 1–9, 2015.

[4] A. Amin, A. Adnan, and S. Anwar, "An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and naïve bayes," *Appl. Soft Comput.*, vol. 137, 2023, Art. no. 110103.

[5] A. Amin, "Just-in-time customer churn prediction in the telecommunication sector," *J. Supercomputing*, vol. 76, pp. 3924–3948, 2020.

[6] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "Counting your customers" the easy way: An alternative to the pareto/NBD model," *Marketing Sci.*, vol. 24, no. 2, pp. 275–284, 2005.

[7] A. Amin et al., "Cross-company customer churn prediction in telecommunication: A comparison of data transformation methods," *Int. J. Inf. Manage.*, vol. 46, no. 4, pp. 304–319, 2019.

[8] A. Amin, F. Al-Obeidat, B. Shah, A. Adnan, J. Loo, and S. Anwar, "Customer churn prediction in telecommunication industry using data certainty," *J. Bus. Res.*, vol. 94, pp. 290–301, 2019.

[9] M. Khajvand and M. J. Tarokh, "Estimating customer future value of different customer segments based on adapted RFM model in retail banking context," *Procedia Comput. Sci.*, vol. 3, pp. 1327–1332, 2011.

[10] P. Anitha and M. M. Patil, "RFM model for customer purchase behavior using K-means algorithm," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 5, pp. 1785–1792, 2022.

[11] P. S. Fader, B. G. S. Hardie, and K. L. Lee, "Computing P (alive) using the BG/NBD model," 2008. [Online.] Available: http://www.brucehardie.com/notes/021/palive_for_BGNBD.pdf

[12] Y. Beeharry and R. T. Fokone, "Hybrid approach using machine learning algorithms for customers' churn prediction in the telecommunications industry," *Concurrency Comput.: Pract. Experience*, vol. 37, 2021, Art. no. e6627.

[13] A. D. Caigny, K. Coussement, and K. W. D. Bock, "A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees," *Eur. J. Oper. Res.*, vol. 269, no. 2, pp. 760–772, 2018.

[14] A. Dingli, V. Marmara, and N. S. Fournier, "Comparison of deep learning algorithms to predict customer churn within a local retail industry," *Int. J. Mach. Learn. Comput.*, vol. 7, no. 5, pp. 128–132, 2017.

[15] S. Wu, W.-C. Yau, T.-S. Ong, and S.-C. Chong, "Integrated churn prediction and customer segmentation framework for telco business," *IEEE Access*, vol. 9, pp. 62118–62136, 2021.

[16] W. Qadadeh and S. Abdallah, "Customers segmentation in the insurance company (TIC) dataset," *Procedia Comput. Sci.*, vol. 144, pp. 277–290, 2018.

[17] S. H. Han, S. X. Lu, and S. C. H. Leung, "Segmentation of telecom customers based on customer value by decision tree model," *Expert Syst. Appl.*, vol. 39, no. 4, pp. 3964–3973, 2012.

[18] S. Peker, A. Koçyiğit, and P. Eren, "A hybrid approach for predicting customers' individual purchase behavior," *Kybernetes*, vol. 46, no. 10, pp. 1614–1631, 2017.

[19] "Online retail II UCI," 2021. Accessed: Aug. 01, 2021. [Online.] Available: https://www.kaggle.com/datasets/mashlyn/online-retail-ii-uci

[20] "Pakistan's largest E-commerce dataset," 2021. Accessed: Aug. 01, 2021 [Online]. Available: https://www.kaggle.com/datasets/zusmani/pakistans-largest-ecommerce-dataset

[21] G. J. Cullinan, "Picking them by their batting averages' recency-frequency-monetary method of controlling circulation," in *Manual Release*, vol. 2103. 1977.

[22] F. R. Dwyer, "Customer lifetime valuation to support marketing decision making," *J. Direct Marketing*, vol. 3, no. 4, pp. 8–15, 1989.

[23] P. D. Berger and N. I. Nasr, "Customer lifetime value: Marketing models and applications," *J. Interactive Marketing*, vol. 12, no. 1, pp. 17–30, 1998.

[24] E. P. Batislam, M. Denizel, and A. Filiztekin, "Empirical validation and comparison of models for customer base analysis," *Int. J. Res. Marketing*, vol. 24, no. 3, pp. 201–209, 2007.

[25] P. S. Fader and B. G. S. Hardie, "The gamma-gamma model of monetary value," 2013. [Online]. Available: http://www.brucehardie.com/notes/025/gamma_gamma.pdf

[26] S.-M. Xie, "Comparative models in customer base analysis: Parametric model and observation-driven model," *J. Bus. Econ. Manage.*, vol. 21, pp. 1731–1751, 2020.

[27] P. Chou, H. H.-C. Chuang, Y.-C. Chou, and T.-P. Liang, "Predictive analytics for customer repurchase: Interdisciplinary integration of buy till you die modeling and machine learning," *Eur. J. Oper. Res.*, vol. 296, no. 2, pp. 635–651, 2022.

[28] D. Sculley, "Web-scale k-means clustering," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 1177–1178.

[29] R. E. Wright, "Logistic regression," in *Reading and Understanding Multivariate Statistics*, L. G. Grimm and P. R. Yarnold, Eds. Washington DC, USA: Amer. Psychol. Assoc., 1995, pp. 217–244.

[30] M. Pal, "Random forest classifier for remote sensing classification," *Int. J. Remote Sens.*, vol. 26, no. 1, pp. 217–222, 2005.

[31] T. F. Chan, G. H. Golub, and R. J. LeVeque, "Updating formulae and a pairwise algorithm for computing sample variances," in *Proc. COMPSTAT 5th Symp. Toulouse, Part I: Proc. Comput. Statist.*, 1982, pp. 30–41.

[32] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Proc. Annu. Conf. Neural Inf. Proccess. Syst.*, 2017, pp. 3146–3154.

[33] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, 1996.

[34] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining*, 2016, pp. 785–794.

[35] S. Balakrishnama and A. Ganapathiraju, "Linear discriminant analysis-a brief tutorial," *Inst. Signal Inf. Process.*, vol. 18, pp. 1–8, 1998.

[36] D. E. Hilt and D. W. Seegrist, "Ridge, a computer program for calculating ridge regression estimates," Dept. Agriculture, Forest Service, Northeastern Forest Experiment Station, Upper Darby, PA, USA, Tech. Rep. NE-236, 1977.

[37] "Classification on imbalanced data—tensorflow core," 2022. [Online]. Available: https://stats.stackexchange.com/questions/244630/difference-between-sample-weight-and-class-weight-randomforest-classifier

[38] "Imbalanced-learn documentation," 2023. Accessed: Jul. 8, 2023. [Online]. Available: https://imbalanced-learn.org/stable/

[39] "BalancedBaggingClassifier," 2023. Accessed: 2023. [Online]. Available: https://imbalanced-learn.org/stable/references/generated/imblearn.ensemble.BalancedBaggingClassifier.html
[40] "Classification on imbalanced data—tensorflow core," 2021. Accessed: Aug. 01, 2021. [Online]. Available: https://www.tensorflow.org/tutorials/structured_data/imbalanced_data
[41] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.

NIEN-TING LEE received the bachelor's degree in computer science and information engineering from Chung Hua University, Hsinchu, Taiwan, in 2020, and the master's degree in electrical engineering from Yuan Ze University, Taoyuan City, Taiwan, in 2022. He is currently a dedicated Engineer with the AI Center of Far Eastern Memorial Hospital, a position he has held from 2022 to 2023. His research interests include artificial intelligence technologies and data analysis.

HAU-CHEN (MIKE) LEE is currently a digital banking strategy Consultant with Far Eastern Group Retail Planning Center, Taiwan. He held senior executive positions in various industries, including as a mobile telecommunication operator and an ETC (electronic toll collection) operator. His research interests include AI/machine learning applications in various industries including behavior modeling for retail customers, air pollution forecasting, crime scene forensic analysis, and IoT. He also serves on the board of listed companies in Taiwan.

JOSEPH HSIN is currently a freshman at Purdue University, West Lafayette, IN, USA, working toward a degree in data science. His research interests include machine learning, natural language processing, and data analysis.

SHIH-HAU FANG (Senior Member, IEEE) is currently a Distinguished Professor with the Department of Electrical Engineering, Director of Innovation Center for AI Applications, and Vice President for Research Development, Yuan Ze University (YZU), Taoyuan City, Taiwan. He has authored or coauthored hundreds of technical journal and conference papers in his research field, which includes indoor positioning, mobile computing, machine learning, and signal processing. Prof. Fang was the recipient of several awards for his research work, including the Young Scholar Research Award (YZU, 2012), Project for Excellent Junior Research Investigators (MOST, 2013), Outstanding Young Electrical Engineer Award (Chinese Institute of Electrical Engineering, 2017), Outstanding Research Award (YZU, 2018), Best Synergy Award (Far Eastern Group, 2018), Future Technology Award (MOST, 2019), National Innovation Award (RBMP, 2019), Y.Z. Outstanding Professor (Y.Z. Hsu Science and Technology Memorial Foundation, 2019), Outstanding Electrical Professor (Chinese Institute of Electrical Engineering, 2021), and Y.Z. Chair Professor (Y.Z. Hsu Science and Technology Memorial Foundation, 2021).