



Segmentation-based Monthly and Weekly Revenue Prediction using K-Means Clustering and XGBoost

Hla Pyae Hein¹, Lai Yee Win Lwin²

Faculty of Information Science
University of Information Technology (UIT), Yangon, Myanmar
hlapyaehein@uit.edu.mm, laiyeewinlwin@uit.edu.mm

ABSTRACT:

There exists substantial fluctuation in retail transaction data among customers, items, regions, and seasons, complicating reliable revenue predictions for Small and Medium-sized Enterprises (SMEs). To address these issues, this research introduces a forecasting framework based on segmentation that integrates K-Means clustering with Local XGBoost for predicting retail revenue on a weekly and monthly basis. The framework is constructed on the Online Retail II dataset and executes segmentation at the customer, product, and geographic tiers prior to forecasting. A chronological data-processing approach is used to maintain the correct temporal sequence. Weekly and monthly revenue series are then constructed for each cluster, together with forecasting features based on historical lags, rolling statistics, seasonal patterns, and holiday indicators. Separate XGBoost models are trained for each segment, while a Direct XGBoost model without clustering serves as the main benchmark for comparison. Model performance is evaluated using Weighted Absolute Percentage Error (WAPE), including mean, holiday-period, and worst-fold performance. The proposed approach achieved a mean WAPE of 19.22%, compared with 24.60% for Direct XGBoost, and also obtained lower errors in the reported holiday-period and worst-fold evaluations. The findings suggest that segmentation can provide useful predictive structure by allowing forecasting models to learn more localized revenue behavior. However, periods of strong seasonal variation remain more difficult to predict, indicating opportunities for further improvement in retail forecasting under volatile conditions.

Keywords: *K-Means clustering, XGBoost, retail revenue forecasting, customer segmentation, time-series forecasting, SMEs.*

1. Introduction

Small and Medium-sized Enterprises (SMEs) are crucial in the retail industry. Specific revenue forecasting can help SMEs in strategic planning, inventory management, allocation of resources, and financial decision-making. Nevertheless, retail income is affected by various purchasing behaviours, product characteristics, geographical conditions, and temporal fluctuations.

1.1 Segmentation-Based Revenue Prediction

Segmentation allows an approach for identifying clusters with similar characteristics prior to revenue forecasting. This research uses K-Means clustering at three different levels, namely customer, product and region. Customer clustering identifies groups of customers with similar buying patterns, product clustering identifies the differences among certain product groups, and regional clustering identifies the differences in revenue trends among different geographic regions. Following the process of segmentation, XGBoost acts as the core predictive model for each level of segmentation. Time series and historical attributes are integrated into the forecasting methodology to produce revenue predictions on a weekly and monthly basis.

1.2 Research Aim

The main aim of this research is to investigate whether segmentation can improve the accuracy of XGBoost-based revenue prediction for retail SMEs. The proposed approach combines *K-Means clustering and XGBoost* at the customer, product, and regional levels and compares the resulting prediction performance with *Direct XGBoost*, where XGBoost is applied without prior segmentation. Other direct prediction models are also used as benchmarks. This work compares these approaches to evaluate whether segmentation-based modelling can provide more accurate monthly and weekly revenue forecasts than directly applying the prediction models to unsegmented retail data.

2. Background

2.1 Customer, Product and Region Segmentation

Segmentation is used to divide retail data into groups with similar characteristics. In this study, K-Means clustering is applied at three levels: customer, product, and region. Such segmentations are useful for identifying different purchasing behaviours, product features and regional revenue trends that could affect revenue prediction.

2.2 K-Means Clustering

K-Means is an unsupervised machine learning algorithm that groups similar data points into clusters based on their features. The algorithm iteratively assigns data points to the nearest cluster centroid and updates the centroids. In this work, we select K-Means as the main clustering technique to discover meaningful segments in retail data.

2.3 XGBoost for Revenue Prediction

XGBoost is a supervised machine learning algorithm based on gradient-boosted decision trees and is widely used for regression and prediction tasks. In this research, XGBoost is used to predict *weekly and monthly revenue* after customer, product, and region segmentation. The segmentation-based results are then compared with direct prediction approaches to evaluate whether segmentation improves prediction accuracy.

3. Research Gap

Existing approaches for retail revenue prediction rely on machine learning algorithms on the entire data set without considering the heterogeneity of customers, products, and regions. However, these groups may show different purchasing habits and revenue patterns that may influence the prediction performance. This study solves this gap by combining segmentation with XGBoost and investigating whether customer, product, and regional segmentation can improve the accuracy of weekly and monthly revenue prediction in comparison with direct prediction methods.

4. Objectives of the Study

1. Clustering K-Means on retail data at customer, product and region level.
2. To develop XGBoost-based models for weekly and monthly revenue prediction.
3. To compare the segmentation-based XGBoost with the Direct XGBoost and the other direct prediction models.
4. To determine the effect of segmentation on the accuracy of revenue forecasting for retail SME

5. Literature Review

5.1 Retail Revenue Forecasting

Previous studies have used statistical and machine learning algorithms to forecast sales and time-series values. Goverdhan et al. (2022) compared ARIMA, XGBoost and LSTM and discovered that XGBoost had a good performance in their forecasting experiments. Their study, however, was limited to financial time series data and did not take into account segmentation before prediction.

5.2 Segmentation-Based Forecasting

Hwang et al. (2025) applied K-Means clustering to classify products into groups with similar sales trends and developed a forecasting model for each group. Their results suggest that more homogeneous groups lead to better forecasting performance. Astakhova et al. (2015) also investigated the use of K-Means clustering for grouped time-series forecasting. Ji et al. (2019) proposed a clustering-based XGBoost approach for sales forecasting in a cross-border e-commerce enterprise, where forecasting models were developed for different product clusters. Similarly, Tang et al. (2024) predicted electricity demand using customer classification, K-Means, and XGBoost. Seyedan et al. (2022) also proposed a cluster-based demand forecasting approach using Bayesian model averaging.

5.3 Comparison with the Proposed Study

The reviewed studies show that clustering and machine learning can be combined for forecasting problems. However, the present study examines segmentation at three retail levels: customer, product, and region. It then applies XGBoost to predict weekly and monthly revenue and compares the segmentation-based results with Direct XGBoost and other benchmark models. This comparison allows the study to evaluate whether segmentation provides an improvement over prediction without prior clustering.

5.4 Summary Comparison

Table 1 - Comparison of Previous Studies and Proposed Study

Aspect	Previous Studies	Proposed Study
Application	Sales, demand, electricity, and financial forecasting	Retail revenue prediction
Segmentation	Usually, one grouping perspective	Customer, product, and region
Clustering	K-Means and related methods	K-Means
Prediction	Statistical, ML, or ensemble models	XGBoost
Comparison	Model-specific comparisons	Segmentation-based XGBoost vs. Direct XGBoost and other benchmarks
Prediction Level	Application-dependent	Weekly and monthly revenue

6. Methodological Framework

This section presents the dataset, segmentation methods, forecasting model, and overall workflow used in the proposed retail revenue prediction framework.

6.1 Dataset

The data that has been used in this research is Online Retail II dataset, which can be accessed from the UCI Machine Learning Repository (Chen, 2019). This dataset contains transaction details of an online retailer in the UK and consists of fields like Invoice id, StockCode, Description, Quantity, Invoice Date, Unit Price, Customer id, and Country. Transaction data spans a period of roughly two years. Revenue for a valid transaction is determined by multiplying the purchased quantity by the respective unit price (Revenue = quantity x unit price). The dataset gives all necessary customer, product, geographical, and temporal information to perform segmentation and predict revenue.

6.2 K-Means Clustering for Retail Segmentation

K-Means clustering algorithm (MacQueen, 1967) is employed to classify the retail data into clusters that have common characteristics. Clustering of retail data is done independently for customer data, product data, and regional data. Different segmentation levels signify different sources of variability in retail revenue and are employed to build the cluster level revenue time series.

6.2.1 Mathematical Formulation of K-Means Clustering

The objective of K-Means clustering algorithm is to group observations such that the sum of distances of all data points from their respective cluster centroid is minimize. The objective function can be written as:

$$J = \sum_{j=1}^K \sum_{x_i \in C_j} \|x_i - \mu_j\|^2 \quad (1)$$

where K is the number of clusters, C_j denotes the jth cluster, x_i is an observation, and μ_j is the centroid of cluster C_j . The centroid is obtained using the equation:

$$\mu_j = \frac{1}{|C_j|} \sum_{x_i \in C_j} x_i \quad (2)$$

Each data point is allocated to the cluster whose centroid is closest by measuring the Euclidean distance:

$$c_i = \arg \min_j \|x_i - \mu_j\|^2 \quad (3)$$

This process continues until the members of each cluster do not change. In this research, K-Means is used separately on the data sets containing customers, products, and regions with K=3, K=5, and K=3, respectively.

6.3 XGBoost for Revenue Prediction

After segmentation is done, XGBoost (Chen & Guestrin, 2016) will be applied as the primary forecast model. The historical revenue data along with time-based features will serve as input for predicting the future revenue of each cluster. The segmented models will be evaluated against Direct XGBoost and other benchmarks to assess whether previous segmentation helped with the forecast.

6.4 Proposed Segmentation and Forecasting Framework

6.4.1 Data Pre-processing

Data pre-processing of transaction data is done prior to clustering and forecasting. All transactions that have been cancelled and transactions that have a non-positive quantity and unit price value are excluded. The revenue of each transaction is computed and those transactions with no registered customers are kept as guest transactions.

6.4.2 Customer, Product, and Region Segmentation

Customer features, product features, and regional features are derived from the cleaned transactions to be used in K-Means clustering. K-Means is applied at three levels: customer (K=3), product (K=5), and region (K=3). The scaler and centroids are fit on the training period only. Later records are assigned to the nearest centroid so that future data do not define the clusters. The resulting cluster identifiers are then attached to the transaction records so that revenue can be analyzed separately for each segment.

6.4.3 Weekly and Monthly Revenue and Forecast Feature Generation

After cluster assignment, transaction revenue is aggregated into weekly and monthly cluster-level time series. Historical lag values, rolling statistics, and calendar-related variables are then generated to represent recent revenue behavior and temporal patterns used by the forecasting model.

6.5 Forecasting Model Formulation

The proposed forecasting process uses historical weekly revenue and monthly to estimate revenue for the following period. The mathematical formulation of the forecasting target and inverse transformation are presented in this section to describe how predictions are generated and assessed. The local XGBoost models are not trained on raw weekly revenue. To reduce scale differences across clusters and to focus the learner on period-to-period variation, the response variable is defined as the first difference of log-transformed revenue:

$$z_{c,t} = \log(1 + y_{c,t}) - \log(1 + y_{c,t-1}) \quad (1)$$

where $y_{c,t}$ denotes observed weekly revenue of cluster c in week t , and $y_{c,t-1}$ is the corresponding lag. XGBoost estimates $\hat{z}_{c,t}$. Because evaluation and managerial interpretation require forecasts on the original monetary scale, the prediction is inverted as:

$$\hat{y}_{c,t} = \exp(\log(1 + y_{c,t-1}) + \hat{z}_{c,t}) - 1 \quad (2)$$

All reported error measures, including WAPE, are computed from $\hat{y}_{c,t}$ and $y_{c,t}$ not from the intermediate log-difference $\hat{z}_{c,t}$.

6.6 Experimental Protocol and Evaluation

Forecasting uses a chronological split of weekly observations into training, validation, and test sets (70%, 15%, and 15%, respectively). Model configuration and selection are performed using the training and validation periods, while the final test period is kept separate from the model-selection process. The proposed Local XGBoost approach is compared with Direct XGBoost, which forecasts the unsegmented weekly revenue series. Model comparison is also performed using rolling-origin evaluation on pre-test observations. The main forecasting metric is WAPE, where lower values indicate better predictive performance.

6.7 Proposed System Design Architecture

The complete proposed system follows the sequence:

Proposed System Design

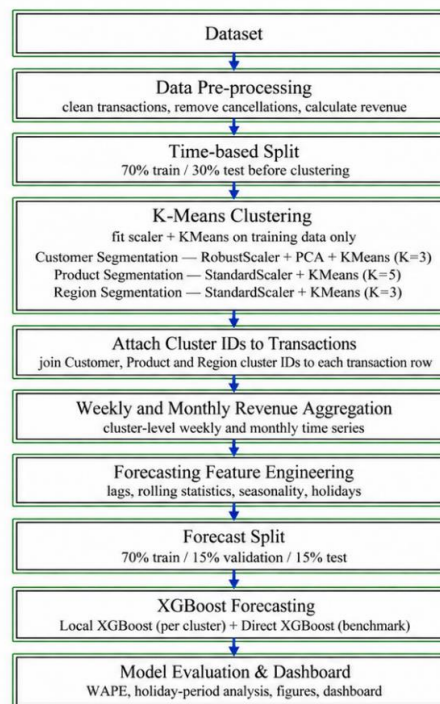


Figure 1: Proposed System Design for Segmentation-based Monthly and Weekly Revenue Prediction

6.8 Flowchart-Based Process Explanation

The proposed workflow starts from the dataset phase, where the transactional data of the Online Retail II dataset are used as the main input for the research. This raw data includes the customer, product, regional, and time data needed for both segmentation and forecasting.

Next, data pre-processing is carried out to clean the transaction data prior to analysis. Any invalid or cancelled transactions are removed, and the revenue is calculated based on quantity and price of each transaction. This will ensure that the input data is fit for clustering and revenue prediction.

Following pre-processing, a time split technique will be applied prior to clustering. The split will be conducted in a way that ensures that the clustering is done based on past data observations. This will ensure that temporal ordering is kept and minimize the chance of using future data to conduct the clustering process.

The processed data will then be passed into K-means Clustering algorithm where segmentation will be done in three different levels, which include customers, products and regions. In this process, different algorithms of clustering will be used to form clusters from similar characteristics of purchasing behaviour, product attributes, and regional revenues. The resulting cluster labels will then be used as the group label in the forecasting phase.

In this attaching cluster IDs to transaction, the allocated customer, product and region clusters IDs will be attached back to the transaction data set.

Next, the transactions are transformed through weekly and monthly revenue aggregation, where revenue is summarized into cluster-level weekly and monthly time series. These series form the main forecasting input because they capture the temporal behaviour of revenue within each segment.

Next step is the forecasting feature engineering process that involves constructing the historical and time-based predictor variables. These include the lagged variables, rolling statistics, seasonality and holiday variables that capture the recent behaviour of the revenue and periodic time patterns.

After the feature engineering process, there is a forecast split. The forecasting data is segmented into training, validation and test periods in a chronological order. The split will be used for the training of the model, model selection and evaluation of the performance of the model.

Next, the segmented forecasting data will be inputted into the XGBoost forecasting process. In the XGBoost forecasting process, the prediction of revenue for future is generated for each segment through the use of the XGBoost models. The forecast framework will be developed in order to model the nonlinear relationships between historical predictor variables and future revenue values.

Finally, the process ends with the model evaluation and dashboards. Selected metrics will be used in evaluating the predicted revenue against the actual data. The results will then be visualized using plots and dashboards.

6.9 Architectural Significance

Through the implementation of the process in a serial manner, the architectural structure creates a relationship among the elements of retail transactions, behavioural segmentation, generation of time series features, and revenue prediction. Through the use of the process structure, the impact of segmentation on the predictive performance can be determined.

7. Results and Discussion

7.1 Customer Cluster Forecasting Performance

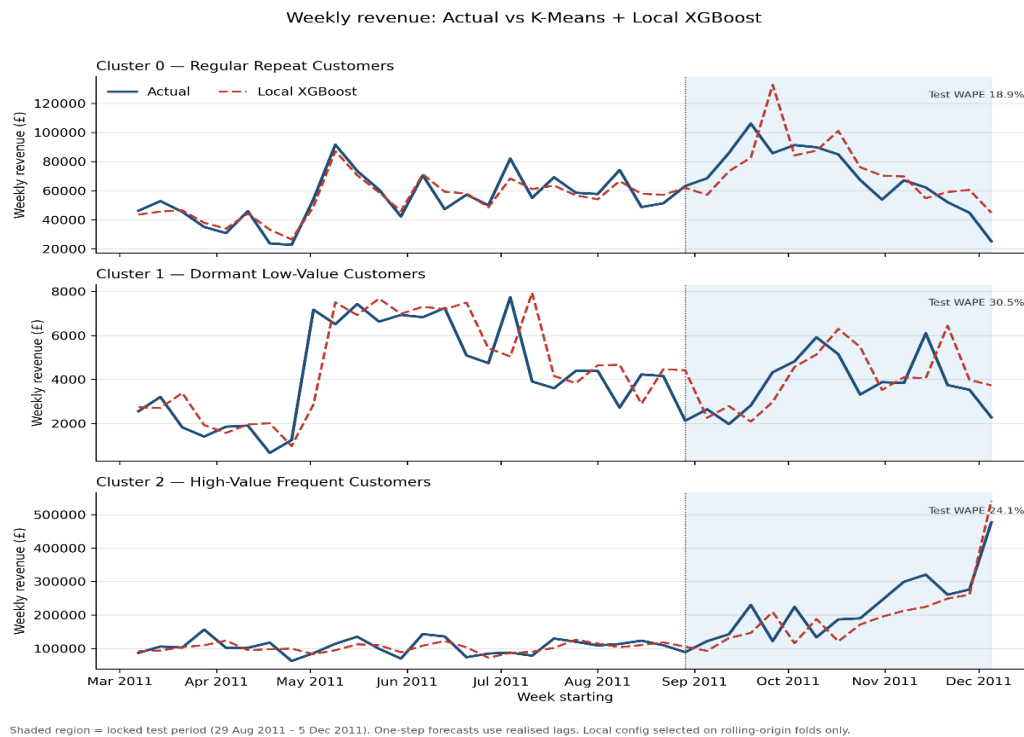


Figure 2: Actual versus Predicted Weekly Revenue for Customer Clusters Using K-Means + Local XGBoost

Figure 2 shows the actual and predicted weekly revenue for the three customer clusters using Local XGBoost. The model achieved test WAPE values of 18.9%, 30.5%, and 24.1% for Clusters 0, 1, and 2, respectively. Overall, the predictions follow the main weekly revenue patterns, although larger errors occur during sudden changes and highly variable periods. For monthly revenue forecasting, the customer clusters achieved WAPE values of 10.5%, 22.9%, and 31.7% for Clusters 0, 1, and 2, respectively. Similar actual-versus-predicted patterns were also observed for the product and regional segmentation levels.

7.2 Forecasting Performance Comparison

Figure 3: WAPE Comparison of Proposed Local XGBoost and Direct XGBoost

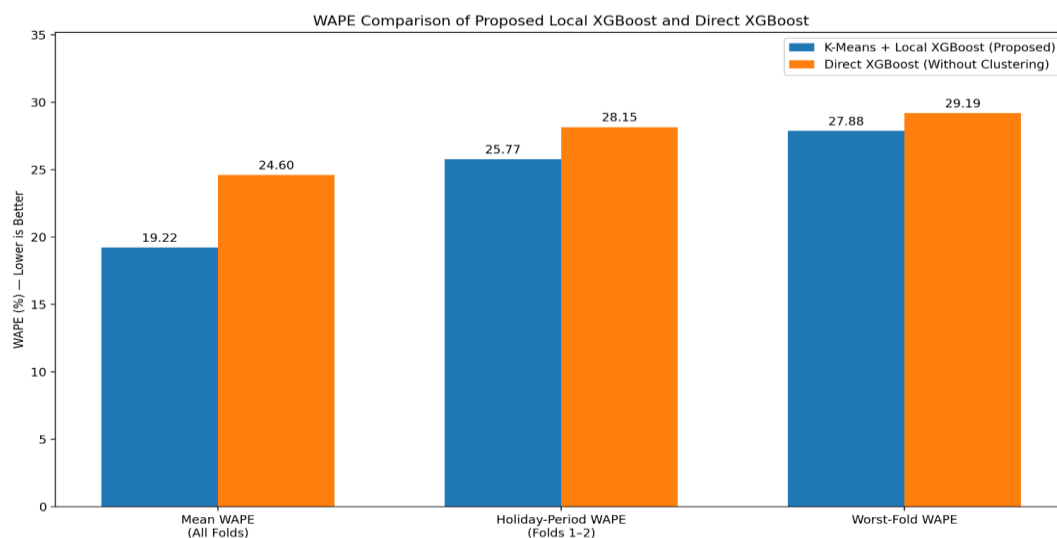


Figure 3 compares the proposed segmentation-based model and Direct XGBoost using mean WAPE across all folds, holiday-period WAPE, and worst-fold WAPE.

The proposed model performed better than Direct XGBoost with a mean WAPE of 19.22% as compared to 24.60%. The proposed model obtained a WAPE of 25.77% in the holiday period folds against Direct XGBoost with 28.15%. The proposed methodology obtained 27.88% for the worst-performing fold compared to 29.19% for Direct XGBoost.

These results show that the proposed segmentation-based approach has resulted in lower forecasting errors in all three tested conditions. The improvement in mean WAPE indicates that the decomposition of retail data into behavioral segments prior to forecasting can provide useful predictive structure. However, both models showed higher error rates during the holiday period and worst-performing fold, suggesting that unusual or highly variable revenue periods are still harder to predict.

7.3 Key Findings

Based on the forecast evaluation, the following results have been established:

1. The proposed K-Means+ Local XGBoost model achieved lower WAPE than Direct XGBoost across the reported evaluation conditions.
2. The minimum error was achieved for the mean performance across all folds with a WAPE of 19.22%.
3. The forecasting errors were increased in holiday and difficult forecasting periods in both models
4. The results indicate that segmentation can support XGBoost forecasting by allowing separate models to capture cluster-specific revenue behaviour.

7.4 Overall Model Performance

Overall, the experimental results indicate that the proposed segmentation-based forecasting framework performs better than applying XGBoost directly to unsegmented revenue data under the reported evaluation conditions. The results support the use of customer behavioral segmentation followed by Local XGBoost forecasting, while also showing that seasonal and unusually volatile periods continue to present forecasting challenges.

8. Conclusion

In this study, a segmentation-based framework combining K-Means clustering and XGBoost is proposed for weekly and monthly retail revenue prediction. Different sources of variation in the retail activity are included in the forecasting process, such as the customer, product and regional segmentation, and historical and time-related features.

The experimental comparison demonstrates that the proposed K-Means + Local XGBoost approach achieved lower WAPE compared to Direct XGBoost under the reported evaluation conditions. However, forecasting errors increased during holiday periods and the most difficult forecasting fold, suggesting that unusual and highly variable revenue patterns remain a challenge. In general, the results indicate that segmentation is a useful preprocessing step to capture differences in retail behavior prior to revenue prediction.

The proposed framework can help SMEs by providing more structured revenue forecasts for planning, inventory management and resource allocation. Future research could focus on the testing of the proposed framework on other retail data sets, incorporating other factors like promotional activities and economic indicators, as well as exploring other forecasting models to improve its performance in peak seasons.

REFERENCES

1. D. Chen, "Online Retail II," UCI Machine Learning Repository, 2019. DOI: 10.24432/C5CG6D.
2. J. MacQueen, "Some methods for classification and analysis of multivariate observations," Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, pp. 281–297, 1967.
3. T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 785–794, 2016.
4. N. N. Astakhova, L. A. Demidova, and E. V. Nikulchev, "Forecasting Method for Grouped Time Series with the Use of k-Means Algorithm," Applied Mathematical Sciences, vol. 9, no. 97, pp. 4813–4830, 2015. DOI: 10.12988/ams.2015.55391.
5. Z. Tang, Y. Jiao, and M. Yuan, "RFM user value tags and XGBoost algorithm for analyzing electricity customer demand data," Systems and Soft Computing, vol. 6, 200098, 2024. DOI: 10.1016/j.sasc.2024.200098.
6. S. Hwang, Y. Lee, B.-K. Jeon, and S. H. Oh, "Sales Forecasting for New Products Using Homogeneity-Based Clustering and Ensemble Method," Electronics, vol. 14, no. 3, p. 520, 2025. DOI: 10.3390/electronics14030520.
7. M. Seyedan, F. Mafakheri, and C. Wang, "Cluster-based demand forecasting using Bayesian model averaging: An ensemble learning approach," Decision Analytics Journal, vol. 3, 100033, 2022. DOI: 10.1016/j.dajour.2022.100033.
8. G. Goverdhan, S. Khare, and M. R., "Time Series Prediction: Comparative Study of ML Models in the Stock Market," Research Article, 2022. DOI: 10.21203/rs.3.rs-2183122/v1.
9. S. Ji, X. Wang, W. Zhao, and D. Guo, "An Application of a Three-Stage XGBoost-Based Model to Sales Forecasting of a Cross-Border E-Commerce Enterprise," Mathematical Problems in Engineering, vol. 2019, Article ID 8503252, 2019. DOI: 10.1155/2019/8503252.