**Eurasia Journal of Science and Technology**
**ISSN: 2663-1016**
**DOI: https://doi.org/10.61784/ejst3071**

# CUSTOMER SEGMENTATION AND CHURN PREDICTION BASED ON K-MEANS AND RANDOM FOREST: A CASE STUDY OF E-COMMERCE DATA

ZhuoRan Li
*School of Economics, Nanjing University of Finance & Economics, Nanjing 210023, Jiangsu, China.*
*Corresponding Email: wxfcg2021@126.com*

**Abstract:** This study aims to segment customers using the application of the K-means clustering algorithm and predict customer churn using the random forest method. Transactional data were used, including the order date, customer name, region, logistics company, quantity bought, payment amount, and frequency bought. K-means clustering was applied to group customers into segments, while a random forest model was constructed to predict customer churn. K-means clustering could determine four customer segments with different purchasing habits. Random forest model could predict customer churn and could find that attributes such as payment value and region were the most significant to use while determining the probability of churn. Results of this study verify that employing K-means clustering and random forest simultaneously for customer segmentation and customer churn prediction is efficient and assists in obtaining considerable insights for precision marketing.

**Keywords:** K-means; Random forest; Customer segmentation; Churn prediction

## 1 INTRODUCTION

In the current era of digital economy, the expansion of e-commerce has revolutionized the competitive landscape significantly, compelling companies to adopt more advanced strategies in a bid to retain customers and facilitate sustainable growth. E-commerce websites have been accumulating massive amounts of customer data, including purchasing behavior, browsing history, and demographic information. These data contain valuable insights into customer behavior and preference that can lead businesses to effective personalized marketing and service optimization. Two significant activities among customer data analysis applications are customer segmentation and churn prediction. Effective customer segmentation enables businesses to personalize marketing to specific customer segments, thereby optimizing customer satisfaction and loyalty. Conversely, churn prediction can identify the potentially churned customers and help businesses take proactive retention measures. These two processes are the crux of customer relationship management (CRM) for e-commerce.Customer segmentation and churn prediction have attracted widespread attention from academia and the business world. Researchers have worked quite a lot in order to improve and develop methods for these tasks, with a lot of analysis achieved through the help of various data mining techniques, some of which are quite advanced, and also various methods [1,2].

### 1.1 Existing Research on Customer Segmentation

The process of customer segmentation is about splitting groups of customers into different kinds of subgroups, which are based on a variety of shared features or characteristics [3]. These can include things like demographics, behavior patterns, and individual preferences [4]. In the past, traditional ways of segmentation often depended on using simple demographic data. However, more recently, with the progress of machine learning, there have been techniques introduced that are much more complex. For instance, K-means clustering is commonly used because it is relatively simple and also effective in managing large sets of data, which makes it a good choice for many situations. Some studies showed how K-means could be applied to segment customers based on their buying habits and the brands they prefer [5]. This kind of segmentation could lead to insights that are very useful, particularly for marketing that is more targeted. There was also a comparison made between K-means and some other clustering methods, such as DBSCAN and agglomerative clustering [6]. In this comparison, it was found that K-means was generally the best performer in terms of two specific measures: the silhouette score and Davies-Bouldin score, which are used to evaluate clustering quality. Moreover, some of the studies that have been done have also explored how segmentation techniques can be connected with business strategies, in order to design marketing campaigns that are better suited to the specific needs and characteristics of each of the segmented customer groups [7].

### 1.2 Existing Research on Customer Churn Prediction

Customer churn prediction is something that looks at the people who might want to leave or end their connection with the business. It's important for the business because it affects their profits, so a lot of attention has been given to this. Many researchers have tried different ways, using various algorithms, and have focused on how selecting the right features and optimizing models can help to make predictions more accurate. A lot of churn prediction relies on machine learning methods, especially random forests, which are popular because they can handle large sets of data and are quite strong in predicting churn. For example, one study from Deng and Gao made an improved K-means algorithm to split customers into groups, then used random forests to predict who might leave, and they were able to spot churners. Some other research also compared random forests with older algorithms, like decision trees or support vector machines, and found that random forests did a better job at predicting customer churn, outperforming them by a good margin. This shows that random forests are preferred in many cases when trying to figure out who's likely to leave a business.

### 1.3 Significance of Applying K-means and Random Forest in E-commerce Customer Segmentation and Churn Prediction

Despite there being progress in both of these fields, research that connects customer segmentation with churn prediction is still, in a way, exploratory. This is especially true in e-commerce. Most studies up to now have been treating these two areas as separate entities, not really taking advantage of the opportunities that might exist if they were to be combined. In the e-commerce industry, combining K-means clustering with random forest algorithms could have some benefits, multiple benefits, actually. For instance, K-means clustering helps businesses identify customer groups based on things like shopping habits, demographics, and other characteristics. This, in turn, may allow for some personalized marketing strategies, or it might help in creating them. Meanwhile, random forests give a strong structure for predicting if a customer will churn or not by figuring out which factors influence their loyalty and whether they'll stay with the brand. The combined use of these two methods can, possibly, improve how well businesses are able to segment customers and predict their likelihood to churn, thus leading to better ways to optimize strategies around retaining customers. This study will, or aims to, explore how both K-means clustering and random forest algorithms might be applied together for customer segmentation and churn prediction in e-commerce. It hopes that by doing this, we will gain a fuller understanding of customer behavior. This, ultimately, could support e-commerce businesses in developing more efficient strategies for managing relationships with their customers.

## 2 METHODS

### 2.1 Data Collection

The data set utilized in this study includes transaction records on an online shopping website between September 2023 and November 2023, with variables including order date, customer name, region, delivery company, purchase quantity, payment amount, and purchase frequency. Following data cleansing and imputation of missing values, a total of 6,419 records were realized over this timeframe.

### 2.2 Data Preprocessing

Preprocessing is the basic step of any data mining process. The data were cleaned to deal with missing values and outliers. The missing values were replaced with mean for numeric attributes and mode (most frequent value) for categorical attributes. The outliers were detected and dropped using the Z-score method. Normalization on the dataset was used so that all features will be contributing equally towards clustering.

### 2.3 Customer Segmentation Using K-means Clustering

K-means clustering, which is a popular unsupervised learning technique, can be used to divide data into K different groups or clusters [8]. The idea is that data points that fall into the same cluster are supposed to be quite similar to each other, whereas those in different clusters are not so much alike [9].

The process of executing this algorithm seems to be something like intuitive. First, it becomes important, or necessary, to determine the number of clusters, K. This step, the selection of K, is regarded as a rather critical one. Usually, in practice, the elbow method is often chosen to decide the optimal K number. The idea behind this elbow method is, or can be, that the sum of squared errors (SSE) is calculated for different values of K, which involves summing the squared distances from each data point to its corresponding cluster center. It is said that when K is small, and as the value of K increases, the SSE decreases considerably, because having more clusters seems to allow for better fitting of data points. However, if you continue increasing K, at a certain point, the decrease in SSE becomes not so sharp or might even level off. This, at some point, may result in a turning point, resembling an elbow, in the graph that shows the K value along with the SSE on the other axis. This turning point in the graph shows the optimal K value, which is believed to be the best.

Once K is chosen, K data points get selected randomly from the dataset, which may serve as the initial cluster centers for the process that will follow. After this selection, an iterative process begins, where in each iteration, the distances from all data points to the K cluster centers are calculated. Usually, a metric like Euclidean distance is employed to determine how far each data point is from a cluster center. Afterward, based on the calculated distance, each data point is assigned to the closest cluster center, or to the cluster that seems nearest. Once all the points have been assigned, the next step is to compute the mean value of the data points in each cluster, and then this mean serves to update the cluster center position. These steps of calculation and update continue to repeat, with the iterations taking place over and over again, until the cluster centers stop moving much or when the number of iterations reaches a preset limit. When the process reaches this point, the algorithm has reached convergence, meaning that the clustering result is obtained. In this study, the elbow method is indeed used to figure out the best K number, with clustering depending on variables that include purchases, frequency, and amount.

### 2.4 Churn Prediction Based on Random Forest

Random forest is an ensemble decision tree learning algorithm and is typically applied in the field of machine learning. Random forest is especially popular due to its high predictive performance as well as robust performance.

To the algorithm principle, random forest creates numerous decision trees and aggregates the prediction results of the decision trees to obtain the final prediction conclusion [10]. In generating each decision tree, random forest adopts a double randomization process. On one hand, through the method of resampling with replacement from the initial training set, a number of bootstrap samples of equal size to the initial set are constructed [11]. Hence, each bootstrap sample may consist of duplicated data, and about 30% of data in the initial set will never appear in the bootstrap sample. Such data are known as out-of-bag data and can be used for model evaluation. On the other hand, when all nodes in the decision tree are split, not all features are utilized. Instead, a random subset of features is sampled from all features, and the best splitting feature is selected from this subset. This random process provides each decision tree some level of distinctiveness, and hence enhances the ability of the model to generalize.

After the model is constructed, for classification problems, random forest uses the voting approach, i.e., each decision tree predicts the sample by prediction, and finally, the category with the most votes is the prediction result of the random forest; for regression problems, the average method is used, and the prediction values of each decision tree are averaged to obtain the final prediction value.

In this paper, whether the customers purchasing this month will continue buying next month or not is shown as "churn" or "not churn". Customers purchasing next month are shown as "not churn", otherwise as "churn". "Whether churn" is used as the target variable, and other variables (order date, customer name, region, logistics company, payment amount, etc.) are used as input variables [8]. Non-numeric variables (such as order date, customer name, area, logistics company, etc.) are coded before they can be analyzed. Random forest modeling is used to determine the major causal factors of e-commerce customer churn in order to facilitate the prediction and prevention of customer churn.

### 2.5 Model Integration

The model proposed, it integrates two methods, K-means clustering and random forest. The purpose is to better understand customer behavior, and it does this by splitting customers into different segments through clustering. Then, using random forest, predictions about customer churn are made. The combination of these two methods, in theory, it works by taking advantage of the strengths from both of them. This can give insights that are useful for businesses, like those in e-commerce, to target customers that belong to particular segments.

### 2.6 Experimental Setup

In this experiment, the tool used is Python, and it includes libraries like scikit-learn and pandas. The dataset for the random forest model, it is divided into two parts: a training set and a test set. The training set gets 70% of the data, and the test set gets 30%. When training, four-fold cross-validation is also applied. This training happens using 70% of the data, and then the remaining 30% is tested. This method, it ensures that the model is evaluated on unseen data after being trained on a large enough portion.

## 3 RESULTS

### 3.1 Customer Segmentation Based on K-means Clustering

**Table 1** Customer Segmentation

| | center value | | |
|---|---|---|---|
| Cluster types | Purchase quantity | Payment amount | Purchase frequency |
| 1 | 5.09 | 227.73 | 1.01 |
| 2 | 49 | 11298.195 | 49 |
| 3 | 17.67 | 3865.65 | 17.67 |
| 4 | 5.17 | 1208.80 | 5.17 |

For K-means clustering, the customers were grouped into four segments. Segment 1 is of consistent purchasing behavior but with low purchase frequency but relatively high payment amount per purchase. They must have special demand for the products but purchase with low frequency. Their purchasing behavior is rational, and they expect a specific quality of the product and services. These can be approached by the firms with high-cost-performance offerings or bundles of services so that they buy more often and remain loyal. Segment 2 includes high-value customers with high frequency of purchase and high payment amount. They are strongly brand-loyal and are the most important customer base for the company. Companies must prioritize retaining these customers by providing quality after-sales service and personalized attention to enhance their word-of-mouth communication and brand loyalty. Segment 3's payment and purchase frequency are balanced, which means stable consumer behavior. These customers likely have stable demand for products. Companies can target this segment with promotional offers or membership schemes to increase their purchase frequency and spending. Segment 4 is defined by low value of payment and purchase frequency, which are low-expenditure buyers. They are price sensitive and require some type of focused marketing efforts that will improve their participation and consumption levels. Companies can provide low-price promotions or coupons to induce this segment to purchase (See Table 1).

### 3.2 Customer Churn Prediction using Random Forest

The impact of random forest on data classification is measured quantitatively by indicators. The accuracy rate and recall rate are better, the higher they are. During this test, the accuracy rate and recall rate of the cross-validation set and test set were checked simultaneously. It can be observed that the two indicators are relatively high. Since the precision rate and recall rate have an effect on one another, if balance between the two needs to be ensured, then the F1 - measure needs to be utilized. The results of the experiments show that the harmonic mean of the precision rate and recall rate of the F1 - measure is also fairly high. The smaller the AUC value, the better the effect of classification (See Table 2).

**Table 2** Classification Evaluation Indicators for Training and Testing sets

| | accuracy | recall | precision | F1 | AUC |
|---|---|---|---|---|---|
| Cross validation set | 0.944 | 0.944 | 0.959 | 0.944 | 0.884 |
| Test set | 0.9 | 0.9 | 0.911 | 0.886 | 0.956 |

**Table 3** Feature Importance Score for Loss Prediction

| Name | Importance |
|---|---|
| Frency | 0.00% |
| Purchase quantity | 1.30% |
| Actual payment amount by the buyer | 78.10% |
| province | 20.50% |

The model indicates that payment amount is the most important attribute that affects customer churn, followed by customer location, and then purchase quantity and frequency of purchases (See Table 3). Why the payment amount is the most important attribute that affects customer churn is as follows. From the perspective of consumption ability and loyalty, the total payment value is the consumption ability and value contribution of the customer to the store. In general, high cumulative payment value customers indicate that they have high awareness of the store's products or services, have established some consumption habits and loyalty, and are more likely to continue creating value for the store. So, there is a very low possibility of their churn. On the other hand, customers with minimal aggregate payment may still be under the trial phase of the store, or lack too much familiarity with the products and services of the store, thus they will churn frequently. According to the input - output psychological view, from the customer's psychological stand, the more the customers spend within the store, the more they will experience that they spent considerable money and time costs in this store. To achieve the best return of the cost incurred by them, they will continue spending in this shop rather than easily moving to other shops.

The reason why customer location plays such an important role in customer churn is the following. From a perspective of geographical differences in consumption culture, the customers of this type of store in various geographical regions may also have different consumption habits and cultures. The customers in some areas can generally be more eager in demand and preferred for this type of store's products or services, and stronger in consumption desire. Thus, customers of such regions are stable, and churn is moderate. Customers in certain regions will have less demand for such kind of products or services, or other substitutes exist in the local market that are more competitive. Such customers will churn more frequently. From the perspective of logistics and convenience of service, the distance between the customer source area and the store and the convenience of logistics distribution will also affect the customer churn rate. Logistics speed is faster in close areas, customers get goods faster, and after-sales service is more convenient. The more convenient customer experience will encourage them to continue shopping at this store. In remote regions with inconvenient logistics, due to long delivery time, high freight costs, or easy occurrence of logistics problems, customer satisfaction may decrease, thus the probability of churn. From the perspective of regional differences in marketing impacts, the marketing investment of the store and the initial implementation of marketing measures in different regions are different. In a few of the key-marketed locations, the customers are more aware and store favorable, and it is easier to establish a stable customer base. In low-marketing coverage areas, the customers have less knowledge of the store and tend to be impacted by other rival stores and defect.

As for the reason why purchase quantity and purchase frequency have relatively smaller effects on customer churn, it is because the purchase quantity and purchase frequency are subject to various factors, such as the special demand cycle of the customer and temporary changes in the external environment. Reasonably speaking, the amount bought and the buying frequency could only reflect the purchase frequency behavior and customer quantity within a specific time period, but cannot reflect fully the customer's loyalty to the store and long-term consumption intention. It is possible that customers will buy in bulk at a single instance due to incidental requirements or promotional events, or have high purchase frequency within a specified period of time, but this does not mean that they will continue buying. Therefore, purchase quantity and frequency shifts cannot accurately reflect whether customers will churn or not. When compared with the total payment amount and customer source region, their performance for predicting customer churn is rather poor.

## 4 DISCUSSION

The results of this study validate that the combination of the K - means clustering and random forest algorithms is effective in analyzing customer behavior and predicting customer churn. Customer segmentation using K - means clustering enables understanding in - depth of the purchasing habits and preferences of different customer segments. By customer segmentation using different customer segments, companies can develop focused marketing strategies to enhance customer satisfaction and loyalty. The high predictive capability of the random forest model of customer churn signifies the importance of key features such as payment value and customer source region. They have critical contributions towards customer loyalty and identifying prospective churners.

When examining the joint analysis outcomes of the two models, it can be found that customers with large cumulative payment amount during the three-month experiment period not only exhibit high purchase frequency and high purchase frequency but also share a relatively concentrated purchase region source. These customers are brand and store loyal and can provide high stability. By contrast, customers with a low purchase amount not only have a low number of purchases and low purchase frequency but also have a dispersed region source, low stability, and are likely to churn. Especially, low - spending customers and price - sensitive customers in the above clusters suffer from serious churn.

Based on this, companies can take a set of measures to increase sales and enhance customer stability, stickiness, and loyalty based on the combined analysis result of the two models. For customer segments with large payment volume and significant purchasing power, companies have to focus on retaining them by rewarding them and improving the shopping experience, e.g., embracing selective offers and personalized services to enhance their shopping pleasure. Especially for regions where there are a large number of customers, considering that they are most likely to be the source of big customers, it is better to increase the intensity of targeted development and marketing, and incline the promotion of direct - access traffic and advertising investment towards these regions. For price-conscious customers with low frequency of consumption, considering their low loyalty and simplicity of their transfer and churn due to small discounts such as price adjustments and their low requirements of brand and quality adjustments, it is recommended to increase their frequency of consumption and rate of purchase conversion through the use of discounts, promotion activities, and accurate recommendations.

In addition, this study also proposes the potential to explore additional characteristics and more advanced models to refine the accuracy of churn prediction. For example, the introduction of more advanced time-series data as well as the integration of variables like the time of last purchase would provide a greater understanding of customer behavior shift. Furthermore, exploring other machine-learning techniques such as gradient boosting or neural networks could deliver better performance for customer-churn prediction.

## 5 CONCLUSION

The study has been successful in utilizing the K - means clustering and random forest algorithm for churn prediction and customer segmentation in the e - commerce industry. The results highlight the importance of payment value and geographical distribution in churn prediction and provide insightful implications for targeted marketing and customer retention techniques. By identifying different customer segments and forecast customer churn well, firms are able to act proactively in addressing potential customer churn problems and enhance customer loyalty. The study confirms that combining K - means clustering and random forest is an efficient approach for customer segmentation and churn prediction and provides a workable model for big data analysis application in e - commerce sector. The findings of this study are also expected to add to the theoretical richness of customer behavior and provide actionable guidance to companies seeking to improve their data - driven decision - making.

Even though the results are encouraging, there are still some limitations to this study. The data for this study were collected from a particular e-commerce platform, and it may not be generalizable to other settings. The applicability of the analytical methods to other datasets and industries has yet to be further tested. In addition, this study did not take into account any external factors influencing customer behavior, such as market trends and economic conditions, that may potentially impact the results and need to be investigated further.

## COMPETING INTERESTS

The authors have no relevant financial or non-financial interests to disclose.

## REFERENCES

[1] Shweta Pandey, Neeraj Pandey, Deepak Chawla. Market segmentation based on customer experience dimensions extracted from online reviews using data mining. Journal of Consumer Marketing, 2023, 40(7): 854-868. DOI: https://doi.org/10.1108/jcm-10-2022-5654.

[2] Petra Jílková. Customer Behaviour and B2C Client Segmentation in Data-Driven Society. International Advances in Economic Research, 2020, 26(3): 325-326. DOI: https://doi.org/10.1007/s11294-020-09799-9.

[3] Tiffany S, Legendre. Consumer value-based edible insect market segmentation [edible insect market segmentation]. Entomological Research, 2020, 51(1): 55-61. DOI: https://doi.org/10.1111/1748-5967.12490.

[4] Deepak Jaiswal, Vikrant Kaushal, Pankaj Singh, et al. Green market segmentation and consumer profiling: a cluster approach to an emerging consumer market. Benchmarking: An International Journal, 2020, 28(3): 792-812. DOI: https://doi.org/10.1108/bij-05-2020-0247.

[5] Rui Zhao. CVM Model of Customer Purchasing Behavior Based on Clustering Analysis. Proceedings of the 2021 3rd International Conference on Economic Management and Cultural Industry (ICEMCI 2021). 2021. DOI: https://doi.org/10.2991/assehr.k.211209.328.

[6] Safae Bouhout,Youness Oubenaalla, El Habib Nfaoui. Comparative Study of Two Parallel Algorithm K-Means and DBSCAN Clustering on Spark Platform. Advanced Intelligent Systems for Sustainable Development (AI2SD' 2020). AI2SD 2020. Advances in Intelligent Systems and Computing, 2022, 1418: 245-262. DOI: https://doi.org/10.1007/978-3-030-90639-9_20.

[7] Wolfgang Bellotti, Daniela N. Davies, Y H Wang. Improved Multi-index Customer Segmentation Model Research. International journal of smart business and technology, 2021, 9 (2): 49-64. DOI: https://doi.org/10.21742/ijsbt.2021.9.2.04.

[8] Girdhar Gopal Ladha, Ravi Singh Pippal. An efficient distance estimation and centroid selection based on k-means clustering for small and large dataset. International journal of advanced technology and engineering exploration, 2020, 7(73): 234-240. DOI: https://doi.org/10.19101/ijatee.2020.762109.

[9] Xiancheng Xiahou, Yoshio Harada. B2C E-Commerce Customer Churn Prediction Based on K-Means and SVM. Journal of Theoretical and Applied Electronic Commerce Research, 2022, 17(2): 458-475. DOI: https://doi.org/10.3390/jtaer17020024.

[10] Feng Ye. Green Progress of Cross-border E-Commerce Industry Utilizing Random Forest Algorithm and Panel Tobit Model. Applied Artificial Intelligence, 2023, 37(1). DOI: https://doi.org/10.1080/08839514.2023.2219561.

[11] Mengyuan Li. Research on the prediction of e-commerce platform user churn based on Random Forest model. 2022 3rd International Conference on Computer Science and Management Technology (ICCSMT), Shanghai, China, 2022, 34-39. DOI: https://doi.org/10.1109/iccsmt58129.2022.00014.