Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 199 (2022) 1332–1339

Procedia Computer Science

www.elsevier.com/locate/procedia

The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

Customer Churn Prediction in Influencer Commerce: An Application of Decision Trees

Sulim Kim[a], Heeseok Lee[b],*

[a] KAIST, College of Business, 85 Hoegiro Dongdaemoon-gu Seoul, 02455, Republic of Korea
[b] KAIST, Digitial Innovation Research Center, 85 Hoegiro Dongdaemoon-gu Seoul, 02455, Republic of Korea

Abstract

This study aims to predict customer churn in influencer commerce. As influencer commerce is a form of e-commerce, influencers directly sell products by uploading website links on their social media account after they promote products through SNS. The role of influencers is to promote brands and/or products on their social media accounts such as Twitter, Facebook, and Instagram. Recently, this role has expanded as a seller. This study implements the customer churn prediction based on the assumption that influencers have passionate support from their followers. The data collected by the influencer marketing agency in Korea from August 2018 to October 2020 includes the purchase details such as customer information, purchase item, and payment amount. In order to predict the churning customers, we apply the Decision Trees (DT) algorithm by using the computer software program, Rapidminer. Our analysis result shows the maximum prediction accuracy is 90% based on F-measure. This study contributes to customer churn prediction from the perspective of influencers.

Keywords: *Customer churn prediction; e-commerce; social commerce; influencer; influencer commerce*

* Corresponding author. Tel.: +82-2-958-3654; fax: +82-2-958-3599.
E-mail address: [a] sulimkim@kaist.ac.kr, [b]hsl@kaist.ac.kr

10.1016/j.procs.2022.01.169

1. **Introduction**

Customers are the most significant assets for any business[2],[5]. Retaining a loyal customer is six times cheaper than recruiting new customers [3]. Predicting churn can help companies refine their goal of retention strategy to minimize losses and enhance marketing decisions [5]. Companies focus on the churn prediction to deliberate what the customer churning points are and cut off the causes preemptively. In particular, the industries which are advantageous to establish long-term relationships with customers, such as airlines, banking, e-commerce, financial services, mobile applications, and online games have contemplated how to retain the churners and what points customers will leave [2]. In recent, the number of customers has been increased for using e-commerce. However, few studies have implemented the churn analysis because it is difficult to define who the churners are in e-commerce. In this study, we intend to conduct a study on predicting customer churn for a new type of e-commerce based on the sales data generated between influencers and customers.

2. Literature Review

2.1. Customer Churn Prediction

Customer churn prediction research has focused on classifying churners by improving the accuracy of models based on various machine learning techniques for the past 5 years. Most customer churn studies focus on advancing model performance by combining new single and ensemble algorithms [1], [16]. finding a method to derive high correlation variables before making a prediction [2],[5] [6], and using new industry data that never applied the churn prediction in the previous studies[8], [9], [10], [11].

First of all, the studies predicted customer churn analysis by applying various machine learning algorithms with the telecommunication data, such as Discriminant Analysis Decision Trees (CART), k-nearest neighbors, Support Vector Machines (SVM), Logistic Regression, Random Forest, Ada Boosting trees, Stochastic Gradient Boosting, Naïve Bayes (NB), Multi-layer perceptron, and Push-Pull-Mooring (PPM) method [1],[16]. Furthermore, setting a data preparation step is another way to increase the accuracy of prediction. The well-established data preprocessing step is directly linked to “the ultimate success of the classification” [4]. Many studies apply the new methods to preprocess the customer data for churning, such as providing guidelines for dealing with the cross-company data, categorical and continuous variables, the clustered customers [3],[5],[6]. Finally, using additional new data is improved the model performance. Customer churn prediction is the allocation of probability to churn to each customer depending on the predicted relationship between the historical data of the customer and the future behavior of the customer [5]. In other words, the ultimate success of prediction depends on what kind of data has been applied to the model. Many churn studies have tried to use the data from various industries such as B2B e-commerce, grocery store, online/mobile game, healthcare application, and dining review application [8], [9], [10], [11].

Thus, most churn prediction literature concentrates on comparing various algorithms to decrease a classification error. They concentrate on improving the accuracy of prediction by applying a new preprocessing method and comparing various algorithms. Few studies have tried to conduct churn prediction in emerging businesses rather than standardized businesses such as telecommunication and finance. This study aims to investigate the customer churn prediction in a new type of e-commerce in social media.

2.2. Churn Prediction in E-commerce

The advances in e-commerce have given customers many choices for purchasing [14]. E-commerce has increased the risk of churn by making it easy to share information between customers, search products, move from one online shopping mall to another [8]. As a result, only a few research on predicting customer churn of e-commerce have been studied due to the unbalanced data and difficulty in clarifying churn point, and the studies are divided into three main themes: developing a new data preprocessing method to eliminate data imbalance, finding out the optimized machine learning algorithms for e-commerce. Berger & Kompan(2019) conducted a customer cluster using the Reference Model in the data preprocessing stage to solve the data imbalance from the online shopping mall and applied the Support Vector Machine(SVM) to analyze the customer churn prediction [2]. Rachid, Abdellah, Belaid, & Rachid(2018) used k-mean and Length-Recency-Frequency-Monetary(LRFM) to cluster the customers and examined the Decision Tree ensemble for predicting churn customers compared to Single Decision Tree, Artificial Neural Network[14]. Yanfang & Chen (2017) developed EBURM (Electronic Business User Retention User) model based on logistic regression in the data preprocessing stage and figured out 'real' customers with the probability of customer loyalty [19]. They verified the accuracy of customer churn prediction by using the AUC model. Sharma & Aggarwal (2020) compared logistic regression, Random Forest, Artificial Neural Networks, and Recurrent Neural Network with e-commerce data, and confirmed that logistic regression hit the highest accuracy of customer churn prediction [16]. Patil, Deepshika, Mittal, Shetty, Hiremath & Patil (2017) applied the random forest, support vector machine, and XGBoost(XGB) models in the model of customer churn respectively based on the online gift retailer data, so that XGB achieved the highest accuracy, 71% [13]. Raeisi & Sajedi(2020) classified e-commerce data with the customer, product, payment, shipping, and customer service information using Gradient Boosted Tree machine learning algorithms, and K-Nearest Nighbor(KNN), Naïve Bayes, Decision Tree, Random Forest, and Rule Induction, and they verified that the accuracy of churn prediction was 86.9% [15].

Xia & He (2018) used online shopping mall data and verified that the classification accuracy of customer churn prediction by mixing BP Neural Network and Support Vector Machine as an ensemble model is better than the single model [18]. They also analyzed the major characteristics of churn customers based on the RFM model. ZHUANG (2018) clustered customers based on the RFM model from e-commerce data and examined the accuracy of customer churn prediction within XG-Boost, Logistic Regression (LR), Support Vector Machine (SVM), and BP Neural Network (BP) [21].

The definitions of churn customers differ in different studies (See Table 1). According to Raeisi & Sajedi (2020), the definitions of churn customers can depend on the services that each company provides. For example, if a month, three months, or even a year have passed since they used the service, they could be classified as churning customers. In addition, Yanfang & Chen (2017) sorted out as churn customers if they have not a history of accessing the e-commerce platform or online shopping mall within 3 months [17]. Zhuang (2018) defined it as a churn if there was no purchase history within the data collection period [21].

In this study, we predict the customer churn using the sales data generated between influencers and customers in social media. Similar to e-commerce, influencers promote and sell items on Instagram by uploading the postings and URL links on their profile. Therefore, we referred to the previous studies that conducted customer churn prediction by using e-commerce to define what the churning point is. If the customer did not buy more than once from an influencer, then the customer is classified as a churner. On the other hand, if it was more than two times, it was classified as loyal.

Table 1: The Summary of e-commerce Customer Churn Prediction Studies

Article	Data	Definition of Churn	Used Algorithms
Gordini & Veglio (2017)	B2B e-commerce (2013.09 ~ 2014.09)	No purchase transaction less than 1 year	Support Vector Machine based on the AUC
Raeisi & Sajedi(2020)	Online food ordering service in Tehran, Iran	No purchase transaction less than 6 months	Gradient Boosting

Rachid, Abdellah, Belaid, & Rachid(2018)	Online shop (2013.11 ~ 2015.02)	Change customer purchase pattern based on LRFM model	Decision Tree Ensemble
Xia & He(2018)	e-commerce website (2014.01~2014.12)	Intermittent lost: no purchase transaction in a certain period Permanent lost: never log in to the account	BP Neural Network & Support Vector Machine
Yanfang & Chen(2017)	e-commerce (3 months)	No login transaction within the recent three months	Logistic Regression
ZHUANG(2018)	e-commerce platform in China (2018.01 ~ 2018.10)	Never purchase transaction in a certain period	XG-Boost

3. Data & Method

3.1. Data

We used the influencer sales data from the influencer marketing agency in Korea. The data obtained customer purchase and refund histories such as the date, seller(influencer), product name, payment amount, refund amount, the number of purchases, and the number of refunds) from August 2018 to October 2020.

Table 2: Descriptive statistics

Variable	Obs.	Mean	Std.Dev	Min	Max
Influencer	510	-	-	-	-
Customer	100,213	-	-	-	-
The number of purchases	-	1.37	1.07	1	28
The number of refunds	-	0.006	0.08	0	2
Average Payment	-	47,057	39,072	1,000	3,858,000
Average Refund	-	232	4,826	0	287,000

Table 3: Training and test set distribution

Training/Testing Set	Type	Number of observations	Percentage
Training Set	Total	39,781	100%
	Churners	31,672	79.6%
	Non-churners	8,109	20.3%
Test Set	Total	87,420	100%
	Churners	71,737	82%
	Non-churners	15,683	17.9%

3.2. Method

This study aims to implement the specialized customer churn prediction algorithm and churning point for influencer commerce based on the previous literature of e-commerce customer churn prediction. The Decision Trees (DT) is a widely used classification algorithm since it is easy to use with high accuracy[9]. Since the value to be predicted is binary, the decision tree is the optimal machine learning algorithm to use [12]. We decide the churning point based on the number of purchases. In other words, if a customer has made a purchase only once from an influencer, it will be considered a churn. The other cases will be classified as loyal (the number of purchases is more than 2 times).

We used Decision Tree (DT) to predict churning customers, which is frequently used in customer churn with e-commerce data. According to Rachid, Abdellah, Belaid, & Rachid (2018) verified the Decision Tree ensemble model had the highest churn prediction accuracy [14]. The Decision Tree (DT) is an algorithm that is frequently used for classification and prediction. It is sometimes used as an ensemble model rather than a single one because of its low performance and robustness. The working principle of DT is to vote for the most popular class in the final output model [14]. To predict churning customers, we implemented 'Rapidminer' which is a computer software program used in data science.

The evaluation of churn prediction can calculate by using these criteria[9]:

- TP(True Positive): "the number of customers that should be in the churner category and the prediction algorithm has determined their category correctly as churner."
- TN(True Negative): "the number of customers that should be in the non-churner but the algorithm incorrectly categorized them as non-churner."
- False Positive: "the number of customers who are non-churners but the algorithm incorrectly categorized them as churners."
- False Negative: "the number of customers who are churners but the algorithm incorrectly categorized them as non-churner."

"Recall is the ratio of real churner which are correctly identified", and its equation as follows:

$$Recall = \frac{TP}{(TP + FN)}$$

"Precision is the ratio of predicted churners which are correct", and its equation as follows:

$$Precision = \frac{TP}{(TP + FP)}$$

"Accuracy is the number of all the correct predictions", and its equation as follows:

$$Accuracy = \frac{(TP + TN)}{(TP + FP + TN + FN)}$$

"F-measure is the harmonic average of precision and recall", and its equation as follows:

$$F - measure = \frac{2 \times Precision \ \times Recall}{Precision + Recall}$$

4. Results

4.1. Customer Churn Prediction

As shown in table 4, we compared the original data and the predicted result. The number of True Positive (TP) is 71,591. As a result, the number of churners in the original data is predicted as the churner. The number of True Negative (TN) is 113. This implies that the number of non-churners from the original data is the same as the predicted result. Moreover, the number of False Positive (FP) and False Negative (FN) are 15,570 and 146. These are clarified as the wrong prediction results.

Table 4: Customer Churn Prediction Result

Data	Type	Status	Predicted		
			Churner	Non-Churner	Total
Original	Count	Churner	71,591	146	71,737
		Non-Churner	15,570	113	15,683
	Proportion	Churner	99.7%	0.2%	100%
		Non-Churner	99.3%	0.7%	100%

4.2. Evaluation

As shown in Table 5, we evaluated the predicted result by calculating recall, precision, accuracy, and F-measure. F-measure is shown as “the real accuracy level” depending on its “significance and importance” [8]. Thus, the maximum prediction accuracy of the Decision Trees (DT) model is 90% regarding f-measure.

Table 5. Summary of Evaluation

Method	Recall	Precision	Accuracy	F-Measure
Decision Trees	99.7%	82.1%	82%	90%

5. Conclusion

Predicting churn customers is important in e-commerce. This study attempts to predict the customer churn in influencer commerce by using Decision Trees (DT) algorithms. Our analysis result demonstrates the feasibility of our method. A further study will be able to apply various algorithms for enhancing applicability.

References

[1] Al-Mashraie, M., Chung, S. H., and Jeon, H. W. 2020. "Customer Switching Behavior Analysis in the Telecommunication Industry Via Push-Pull-Mooring Framework: A Machine Learning Approach," Computers & Industrial Engineering (144).

[2] Amin, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., and Anwar, S. 2019a. "Customer Churn Prediction in Telecommunication Industry Using Data Certainty," Journal of Business Research (94), pp. 290-301.

[3] Amin, A., Shah, B., Khattak, A. M., Lopes Moreira, F. J., Ali, G., Rocha, A., and Anwar, S. 2019b. "Cross-Company Customer Churn Prediction in Telecommunication: A Comparison of Data Transformation Methods," International Journal of Information Management (46), pp. 304-319.

[4] Berger, P., & Kompan, M. (2019). User modeling for churn prediction in E-commerce. IEEE Intelligent Systems, 34(2), 44-52.

[5] Coussement, K., Lessmann, S., and Verstraeten, G. 2017. "A Comparative Analysis of Data Preparation Algorithms for Customer Churn Prediction: A Case Study in the Telecommunication Industry," Decision Support Systems (95), pp. 27-36.

[6] De Caigny, A., Coussement, K., and De Bock, K. W. 2018. "A New Hybrid Classification Algorithm for Customer Churn Prediction Based on Logistic Regression and Decision Trees," European Journal of Operational Research (269:2), pp. 760-772.

[7] De Caigny, A., Coussement, K., De Bock, K. W., and Lessmann, S. 2020. "Incorporating Textual Information in Customer Churn Prediction Models Based on a Convolutional Neural Network," International Journal of Forecasting (36:4), pp. 1563-1578.

[8] Gordini, N., and Veglio, V. 2017. "Customers Churn Prediction and Marketing Retention Strategies. An Application of Support Vector Machines Based on the Auc Parameter-Selection Technique in B2b E-Commerce Industry," Industrial Marketing Management (62), pp. 100-107.

[9] Khodabandehlou, S., and Zivari Rahman, M. 2017. "Comparison of Supervised Machine Learning Techniques for Customer Churn Prediction Based on Analysis of Customer Behavior," Journal of Systems and Information Technology (19:1/2), pp. 65-93.

[10] Kwon, H., Kim, H. H., An, J., Lee, J.-H., and Park, Y. R. 2021. "Lifelog Data-Based Prediction Model of Digital Health Care App Customer Churn: Retrospective Observational Study," Journal of Medical Internet Research (23:1), p. e22184.

[11] Kwon, Y. D., Chatzopoulos, D., ul Haq, E., Wong, R. C.-W., and Hui, P. 2019. "Geolifecycle," Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies (3:3), pp. 1-29.

[12] Lee, A. S. H., Claudia, N., Zainol, Z., & Chan, K. W. (2019, August). Decision Tree: Customer churn analysis for a loyalty program using a data mining algorithm. In International Conference on Soft Computing in Data Science (pp. 14-27). Springer, Singapore.

[13] Patil, A. P., Deepshika, M. P., Mittal, S., Shetty, S., Hiremath, S. S., & Patil, Y. E. (2017, August). Customer churn prediction for retail business. In 2017 International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS) (pp. 845-851). IEEE.

[14] Rachid, A. D., Abdellah, A., Belaid, B., & Rachid, L. (2018). Clustering prediction techniques in defining and predicting customers defection: The case of e-commerce context. International Journal of Electrical and Computer Engineering, 8(4), 2367.

[15] Raeisi, S., & Sajedi, H. (2020, October). E-Commerce Customer Churn Prediction By Gradient Boosted Trees. In 2020 10th International Conference on Computer and Knowledge Engineering (ICCKE) (pp. 055-059). IEEE.

[16] Sabbeh, S. F. 2018. "Machine-Learning Techniques for Customer Retention: A Comparative Study," International Journal of advanced computer Science and applications (9:2).

[17] Sharma, K., Licsandru, T. C., Gupta, S., Aggarwal, S., & Kanungo, R. (2020). An investigation into corporate trust and its linkages. Journal of Business Research, 117, 806-824.

[18] Xia, G., & He, Q. (2018, March). The Research of Online Shopping Customer Churn Prediction Based on

Integrated Learning. In 2018 International Conference on Mechanical, Electronic, Control and Automation Engineering (MECAE 2018) (pp. 259-267). Atlantis Press.

[19] Yanfang, Q., & Chen, L. (2017, December). Research on E-commerce user churn prediction based on logistic regression. In 2017 IEEE 2nd Information Technology, Networking, Electronic and Automation Control Conference (ITNEC) (pp. 87-91). IEEE.

[20] Zhu, B., Baesens, B., and vanden Broucke, S. K. L. M. 2017. "An Empirical Comparison of Techniques for the Class Imbalance Problem in Churn Prediction," Information Sciences (408), pp. 84-99.

[21] ZHUANG, Y. (2018). Research on E-commerce Customer Churn Prediction Based on Improved Value Model and XG-Boost Algorithm. Management Science and Engineering, 12(3), 51-56.