sciendo International Journal of Mathematics and Computer in Engineering 1(2) (2023) 171–176



International Journal of Mathematics and Computer in Engineering

https://sciendo.com/journal/IJMCE

Original Study

Machine learning based churn analysis for sellers on the e-commerce marketplace

Mehmet Emin Öztürk[1], Akasya Akyüz Tunç[1], Mehmet Fatih Akay[2] †

[1]Trendyol, Data Science, İstanbul, Türkiye

[2]Department of Computer Engineering, Cukurova University, Adana, Türkiye

Communicated by Haci Mehmet Baskonus; Received: 15.06.2023; Accepted: 19.07.2023; Online: 20.07.2023

Abstract

The goal of this study is to develop churn models for sellers on the e-commerce marketplace by using machine learning methods. In order to develop these models, three approaches are applied. The dataset used in this study includes ten features, which are maturity type, maturity interval, city of the seller, total revenue of the seller, total transaction of the seller, sector type of the seller, business type of the seller, sales channel, installment option and discount type. Random Forest (RF) and Logistic Regression (LR) are used for churn analysis in all of the approaches. In the first approach, models are developed without applying preprocessing operations on the dataset. In the second and third approaches, under sampling and oversampling methods are used respectively to balance the data set. By using stratified cross validation on the dataset, F-Scores of the churn models are obtained. These results show that F-Scores were 0.76, 0.71 and 0.92 for the three approaches developed with RF, and 0.84, 0.68 and 0.69 for the three approaches developed with LR, respectively.

Keywords: Churn analysis, machine learning, marketplace.
AMS 2020 codes: 68Txx; 68T07; 68T20; 68T01; 68T09.

1 Introduction

Interest in e-commerce, which is one of the fastest-growing sectors in the world, is increasing day by day. The main reason why users choose e-commerce is that they can shop effortlessly, around the clock, from any location, simply by using an internet connection and with a wide range of options. The term E-commerce marketplace refers to online platforms set up by third parties on the Internet to bring sellers and buyers together. Sellers can thus open stores and upload their products, while consumers rate and buy the corresponding products.

E-commerce companies are constantly trying to attract new customers and increase revenue to ensure their continued existence. However, research has shown that acquiring a new customer is more expensive than retaining an existing one. For this reason, companies spend most of their efforts on customer retention. However, the efforts made with the wrong strategy lead to negative consequences for companies. Assessing the risk of churn is an important task in customer retention.

In the last few years, numerous methods have been used for churn analysis. [1] presented a comprehensive evaluation of the capacity of Recurrent Neural Networks (RNN) and transducers for Customer Attrition Prediction (CCP) using time-varying behavioral characteristics such as novelty, frequency, and monetary value.

†Corresponding author.
Email address: mfakay@cu.edu.tr

sciendo
ISSN 2956-7068
doi:10.2478/ijmce-2023-0013

Open Access.© 2023 Akay et al., published by Sciendo.
This work is licensed under the Creative Commons Attribution alone 4.0 License.

Hybrid approaches combining DNN outputs from traditional CCP models were also evaluated in this context. The obtained results demonstrated that DNNs, particularly RNNs, were appropriate for CCP using time-varying recency, frequency, and monetary measurements. [2] proposed a Naive Bayes (NB) classifier with a Genetic Algorithm (GA) based feature weighting approach for loss estimation. They also compared the proposed approach to the classifier in public datasets, with mean precision of 0.97, 0.97, 0.98, and sensitivity rates of 0.84, 0.94, 0.97. The proposed method yielded F scores of 0.89, 0.96, 0.97, Matthews Correlation Coefficients of 0.89, 0.96, 0.97, and accuracies of 0.95, 0.97, 0.98. [3] presented a hybrid recommendation strategy based on SVM with targeted retention attempts for an e-commerce loss forecast. When the integrated forecasting model was used, the coverage rate, hit rate, removal rate, precision rate, and other metrics increased significantly. [4] described how the organization could create a customer churn model based on mathematical and statistical approaches using customer acquisition, usage, interaction, and call center data, with the goal of minimizing customer churn. This research has aided many organizations and data scientists in developing a customer churn model. [5] proposed a theoretical foundation for customer churn and customer segmentation, as well as the use of supervised machine learning techniques to predict churn. Customers were segmented using K-Means, K-Nearest Neighbors (KNN), Logistic Regression (LR), Decision Tree (DT), and RF. They employed the SVM to forecast customer churn. As a result, the data set achieved an Accuracy of approximately 97% with the RF model, and the average Accuracy of each model performed well after customer segmentation, with LR having the lowest Accuracy (87.27%). [6]performed data preprocessing and feature analysis in the first two stages of churn analysis. In the prediction process, most popular predictive models namely Support Vector Machine (SVM), Decision Trees (DT), Naive Bayes, LR, RF as well as boosting and ensemble techniques have been applied. According to the results, AdaBoost and XGBoost classifier were found to give the highest accuracy. In order to classify the data more accurately, [7] suggested a brand-new hybrid method called Logit Leaf Model (LLM). LLM consists of estimation and segmentation stages. It was seen that different models created by LLM on data segments instead of the whole dataset provided better prediction performance. LLM provided more accurate models than using the independent classification techniques LR and DT. To help telecom providers estimate which customers they are most likely to lose, [8] generated a model. The Area Under the Curve measure was adopted to measure the performance of the model. Additionally, it was intended to incorporate social network analysis elements into the prediction model in order to incorporate customer social networks. The best results in the model were obtained when the eXtreme Gradient Boosting (XGBoost) algorithm was applied. [9] suggested a hypothesis to predict customer churn using deep learning-based neural networks. In order to explain the interactions, this article considered mathematical models and modeling methods were also suggested. For chain training, Python was used by processing user activity datasets. [10] examined six different methods for estimating customer churn for the banking industry. The best results were obtained with Stochastic Boosting. [11] calculated consumer turnover using deep learning on a telecom dataset. It was decided to build a nonlinear classification model using a multilayer neural network. There was research done on the churn model, customer features, support features, use features, and contextual features. [12] aimed to provide a framework that attempts to perform churn analysis. Articial Neural Network (ANN) and DT were used as methods. According to the results obtained, five variables including number of products, acceptance of returned products, discount, delivery time and reward were selected as the best variables. ANN had the highest accuracy with 97.92% accuracy, while DT had the least.

This paper is organized as follows: Section 2 presents material and method. Results and discussion are presented in Section 3. Section 4 concludes the paper.

2 Material and method

2.1 Random forest

The output of various decision trees is combined with the widely used machine learning algorithm RF to produce a single outcome. Its widespread use is motivated by its adaptability and usability because it can solve

classification and regression issues. The working logic is to select the number of trees to be created and each of the trees created to predict the y value for each data point [13]. The hyperparameters used for the RF method are listed in Table 1.

Table 1 Hyperparameter values of RF.

Hyperparameter	Values
Number_of_estimators	100
Criterion	gini
Minimum_samples_split	2
Minimum_samples_leaf	1

2.2 Logistic regression

When the number of dependent variables is two, LR is the proper regression strategy to use. The LR is predictive analysis, as all regression analyses are. Using LR, data can be described and the relationship between a dependent binary variable and one or more independent nominal, ordinal, interval, or ratio-level variables can be explained [14]. The hyperparameters used for the LR method are listed in Table 2.

Table 2 Hyperparameter values of LR.

Hyperparameter	Values
Solver	sag
C	100
Max_iter	1000
Tol	0.0001

To increase the performance of the model, the best values of the hyperparameters of the RF and LR have been found by grid search.

Class imbalance, also known as the unbalanced data problem, is a problem that must be considered in data-driven studies. Most classification algorithms assume that the training data sets are balanced. However, this assumed balanced distribution is often not found in real datasets. One of the classes may be represented by very few instances, while the other class is represented by a large number of instances. In this case, classification problems may occur. For samples with little labeling information, the model is likely to make incorrect predictions for this group because the model is not sufficiently trained. The first method that can be used when working with an imbalanced data set is to adjust the class distributions by resampling the data. These methods are undersampling and oversampling.

2.3 Undersampling

Undersampling is a technique to balance the data set by removing samples that belong to the majority class. It is one of many methods data scientists can employ to draw out more accurate information from datasets that were initially unbalanced. Although it has drawbacks like the loss of potentially crucial information, data scientists nonetheless frequently use and value this expertise [15].

2.4 Oversampling

Oversampling balances the data set by duplicating samples belonging to the minority class. Although it has drawbacks, such as the fabrication of synthetic data, it has become a frequent and crucial skill for data scientists. [16].

3 Results and discussion

Churn models have been developed for sellers in the e-commerce market using machine learning methods. The features in the data set used in the study are maturity type, maturity range, seller's city, seller's total income, seller's total transaction, seller's sector type, seller's business type, sales channel, installment option and discount type. Three different approaches were applied to develop the models. RF and LR were used for churn analysis in all of the approaches. In the first approach, models were developed without applying preprocessing operations on the dataset. In the second and third approaches, undersampling and oversampling methods, respectively, were used to balance the data set. The reason for using three different approaches is to analyze the impact of the preprocessing of the data set on the performance of the models. Cross-validation is a technique used to predict how a model will perform with real data obtained with training data. In this technique, the model is trained with training data while the performance of the model is evaluated using the remaining data (validation data). By using stratified cross-validation on the dataset, F-Scores of the churn models were obtained. Table 3 shows F-Scores of the models developed with three different approaches using RF and LR.

Table 3 F-Score values obtained with three different approaches using RF and LR.

Approach	RF	LR
First Approach	0.76	0.84
Second Approach	0.71	0.68
Third Approach	0.92	0.69

– When all prediction models developed with RF are compared, it has been determined that the highest F-Score value was obtained with the third approach using the oversampling technique.

–When the F-Scores obtained by the RF method were examined, it was seen that the undersampling technique did not make much difference compared to the models created on the unbalanced data set.

–When all prediction models developed with LR are compared, it has been determined that the highest F-Score value was obtained with the first approach which was created without preprocessing.

–When the results obtained with the LR method were examined, it was seen that the oversampling/undersampling techniques significantly reduced the F-Scores of the models compared to the model created on the unbalanced data set.

–When the results of both methods are examined, it was seen that the models developed using the undersampling technique had the lowest F-Score.

–It was determined that the mean F-Scores obtained with the RF method (0.80) was 8.11% higher than that obtained with the LR method (0.74).

4 Conclusion

In this study, churn models are developed for sellers in the e-commerce market by using machine learning based RF and LR methods. Three approaches were used to develop the models. In the first approach, models are developed without preprocessing the dataset. In the second and third approaches, undersampling and oversampling methods, respectively, were used to balance the data set. F-Score was used to assess how effectively the churn models performed. The results showed that the model developed using the RF oversampling technique was the most successful one. It can be concluded that the developed churn model is effective and can be used in practice.

5 Declarations

5.1 Conflict of interest:

Authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

5.2 Funding:

The authors hereby declare that there is no any funding interest.

5.3 Authors Contribution:

M.E.Ö.-Conceptualization, Data Curation, Writing-Review Editing. A.A.T.-Software, Data Curation, Writing-Original Draft. M.F.A.-Supervisior, Methodology, Writing-Original Draft. All authors read and approved the final submitted version of this manuscript.

5.4 Acknowledgement:

The authors sincerely appreciate the reviewers' insightful remarks, which resulted in a significant improvement in the manuscript.

5.5 Data availability statement:

All data that support the findings of this study are included within the article.

References

[1] Mena G., Coussement K., Bock K.W.D., Caigny A.D., Lessmann S., Exploiting time-varying RFM measures for customer churn prediction with deep neural networks, Annals of Operations Research, 1(23), 1-23, 2023.
[2] Amin A., Adnan A., Anwar S., An adaptive learning approach for customer churn prediction in the telecommunication industry using evolutionary computation and Naïve Bayes, Applied Soft Computing, 137(110103), 1-12, 2023.
[3] Shobana J., Gangadhar C., Arora R.K., Renjith P.N., Bamini J., Chincholkar Y.D., E-commerce customer churn prevention using machine learning-based business intelligence strategy, Measurement: Sensors, 27(100728), 1-8, 2023.
[4] Sharma A., Predictive Modelling (Customer Churn) Measurement and Prevention, Authorea, DOI:10.22541/au.167842113.31493988/v1, 2023.
[5] Tran H.D., Le N., Nguyen V.H., Customer churn prediction in the banking sector using machine learning-based classification models, Interdisciplinary Journal of Information Knowledge and Management, 18, 87-105, 2023.
[6] Lalwani P., Mishra M.K., Chadha J.S., Sethi P., Customer churn prediction system: a machine learning approach, Computing, 104(2), 271-294, 2022.
[7] Caigny A.D., Coussement K., Bock K.W.D., A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees, European Journal of Operational Research, 269(2), 760-772, 2018.
[8] Ahmad A.K., Jafar A., Aljoumaa K., Customer churn prediction in telecom using machine learning in big data platform,

Journal of Big Data, 6(28), 1-24, 2019.
[9] Morozov V., Mezentseva O., Kolomiiets A., Proskurin M., Intellectual systems of decision making and problem of computational intelligence predicting customer churn using machine learning in IT startups, International Scientific Conference, 645-664, 2021.
[10] Dias J., Godinho P., Torres P., Machine learning for customer churn prediction in retail banking, Computational Science and Its Applications-ICCSA 2020, Springer, 2020.
[11] Agrawal S., Das A., Gaikwad A., Dhage S., Customer churn prediction modelling based on behavioural patterns analysis using deep learning, IEEE Access, 10, 14529-14547, 2022.
[12] Khodabandehlou S., Rahman M., 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), 65-93, 2017.
[13] Ali J., Khan R., Ahmad N., Maqsood I., Random forests and decision trees, International Journal of Computer Science Issues, 9(5), 272-278, 2012.
[14] Arafa A.A., Radad M., Badawy M., Fishawy N.E., Logistic regression hyperparameter optimization for cancer classification, Menoufia Journal of Electronic Engineering Research, 31(1), 1-8, 2022.
[15] Anand A., Pugalenthi G., Fogel G.B., Suganthan P.N., An approach for classification of highly imbalanced data using weighting and undersampling, Amino Acids, 39(5), 1385-1391, 2010.
[16] Zhu T., Lin Y., Liu Y., Synthetic minority oversampling technique for multiclass imbalance problems, Pattern Recognition, 72, 327-340, 2017.