

Федеральное государственное бюджетное образовательное учреждение  
высшего образования  
«Новосибирский государственный технический университет»

Кафедра иностранных языков технических факультетов

**Реферат**

по дисциплине «Английский язык»

Тема: User Churn Model in E-Commerce Retail

Source: Fridrich, M., & Dostál, P. (2022). User Churn Model in E-Commerce Retail. Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration, 30(1). <https://doi.org/10.46585/sp28031105>

Рецензия: текст соответствует  
структуре учебно-научн.  
моногр. реф.

Выполнил:

Студент: Сухих А.С,

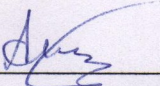
Группа ПММ-21

Проверил:

Преподаватель: Балобанова А.Г.

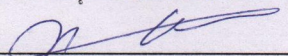
Балл: 9 (82) B+

Оценка отлично



подпись

« 01 » декабря 20 22 г.



подпись

« 01 » декабря 20 22 г.

Новосибирск 2022

# **User Churn Model in E-Commerce Retail**

Martin Fridrich, Petr Dostal

Brno University of Technology, Faculty of Business and Management, Institute  
of Informatics, Czech Republic

The rapid growth of the retail e-commerce created competitive environment, which led organizations to focus on retaining customers. One of the key component of customer retention is predicting customer churn. The authors analyzed several articles dealing with customer churn resulting in the comparison table, that describes research articles, their experiment, modeling algorithms and performance metrics. Modeling pipeline in considered studies

It is said, that despite the fact that many researches already were provided in discussed area, there is no consensus on a customer churn model, concerning both explanatory and explained characteristics. Thus the main goal of this study is to propose a viable user churn model based on a traditional and new set of attributes.

The authors form the model around user-item interactions, so the churn event, which is the dependent variable, is defined as interaction/no interaction with the e-commerce website during next month unlike transactional define of customer churn in other studies. The user model (independent variables) consists of 6 sets: recency (the time since the last transaction), frequency (number of transactions within a period), monetary (total revenue within a period), category and item (customer's preference), date-time (user experience level during the day) and other (time-to-event features, average session length) characteristics.

To ensure transparency and reproducibility authors introduced an open e-commerce retail dataset, which covers the period of 2015/05/09-2015/09/17, consists of 49358 observations, 47 user model vectors and the churn event. The target class distribution is imbalanced with a churn rate of 89%. Further dataset analysis with triangular matrix of Pearson's correlation revealed that the attributes are asymmetric and suffer from outliers.

This dataset is used to build a customer churn model. The first step of the modeling pipeline is Data Processing, which includes imputation of missing values,

omitting the vectors with near-zero variance and adding second-degree polynomial expansion. Further in Feature Extraction step the principal component analysis was applied to project the data into orthonormal 50-dimensional space to mitigate multicollinearity and sparsity of independent variables. Finally, for modeling authors used out-of-the-box implementations of classification algorithms hinged on scikit-learn library. Regularized logistic regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting Machine (GBM) algorithms were used in the study.

Performance benchmark included 20-fold cross-validation scheme, which consists of dividing dataset on 20 data splits and modeling procedure until each split acts like a validation split. For comparing different algorithms authors used Accuracy, F1-score and Area under the receiver operating characteristic Curve metrics.

The research results indicate LR, SVM and GBM algorithms almost on par in Accuracy and F1-score, having LR as the best-performing classifier. Talking about AUC metric GBM achieved the highest value. The null hypothesis in these metrics evaluation wasn't rejected. The remaining solutions did not perform well considering acceptance of the alternative hypothesis.

These best-performing classifiers (LR, SVM and GBM) was used to evaluate feature importance. The Recency set exhibits outstanding qualities with a very high F1; 65–86 % of its elements are identified as relevant; modeling solutions select 33–43 % of the essential characteristics from the Recency set. The second was Frequency group, with a considerable F1; 36–52% relevant and 25–32% of essential characteristics in LR and GBM. Category & item and Date & time characteristics have a moderate F1; these sets suffer from sparsity and is the essential set for SVM-RBF; the latter group shows acceptable relevant elements and is favoured by GBM. Difference in mean point estimates showed that Recency and Frequency surpass other sets in all aspects. Date & time also performed well combined with GBM.

Nowadays rapid growth of e-commerce retail has made the predicting of customer churn a pretty discussible topic. It resulted a big amount of related research literature but no agreement concerning customer churn model were demonstated.

This study proposed a model that can be used in future researches and also aid business decision-making efforts related with churn prediction. Besides this research showed that the Recency and Frequency of sessions of a customer have a very high correlation with customer loyalty. Thus a recession of customers visits may be an indicator of increasing customer churn.

In conclusion the authors said that a customer churn model suitable for e-commerce retail were proposed based on user interactions with the website. The best-performing solutions are LR, SVM and GBM. Recency (period from the last session) and Frequency (average period between the sessions and number of sessions) occurred to be the most important characteristics of customer churn.