SciPap
scientific papers

Scientific Papers of the University of Pardubice, Series D: Faculty of Economics and Administration 2022, 30(1), 1478.

DOI: 10.46585/sp30011478
editorial.upce.cz/SciPap

User Churn Model in E-Commerce Retail

Martin Fridrich

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

Petr Dostál

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

Abstract

In e-commerce retail, maintaining a healthy customer base through retention management is necessary. Churn prediction efforts support the goal of retention and rely upon dependent and independent characteristics. Unfortunately, there does not appear to be a consensus regarding a user churn model. Thus, our goal is to propose a model based on a traditional and new set of attributes and explore its properties using auxiliary evaluation. Individual variable importance is assessed using the best performing modeling pipelines and a permutation procedure. In addition, we estimate the effects on the performance and quality of a feature set using an original technique based on importance ranking and information retrieval. The performance benchmark reveals satisfying pipelines utilizing LR, SVM-RBF, and GBM learners. The solutions rely profoundly on traditional recency and frequency aspects of user behavior. Interestingly, SVM-RBF and GBM exploit the potential of more subtle elements describing user preferences or date-time behavioural patterns. The collected evidence may also aid business decision-making associated with churn prediction efforts, e.g., retention campaign design.

Keywords

User Model, Churn Prediction, Customer Relationship Management, Electronic Commerce, Retail, Machine Learning, Feature Importance, Feature Set Importance

JEL Classification

C60, M31

Introduction

During the last two decades, the unprecedented growth of the retail e-commerce industry has created a competitive environment built upon immense technological advances and a user-centric paradigm (MacKenzie et al., 2013; Morgan, 2018; Terdiman, 2018). It has become essential for organizations to sustain a valuable user/customer base. Churn mitigation is principally motivated by a disparity among the unitary costs of user acquisition and retention (Gronwald, 2017); however, there are additional benefits (Ascarza et al., 2018).

The retention strategy is arranged into short- and long-term pursuits (Mezghani et al., 2012; Ascarza et al., 2018). In the short term, companies aim to capture behavioural and transactional dynamics to identify users at imminent risk of churning and preserve them if profitable. In the long term, firms lean towards enhancing customer satisfaction and loyalty through understanding the drivers of churn. Nevertheless, the success of both pursuits is highly dependent on appropriate representation. Brusilovsky (1996) proposes a user model concept to describe user features such as behavior, goals, knowledge, stereotypes, and preferences linked to practical actions. Its quality can be assessed indirectly with machine learning methods. Unfortunately, there is no clear consensus on a user/customer churn model concerning both explanatory and explained characteristics.

Hence, we aim to propose a user churn model for e-commerce retail and explore its properties using indirect evaluation. The rest of the article is structured as follows: the research literature concerning the user model, modeling, and evaluation is covered in the second section. The user churn model is introduced and reviewed in the third section. The following part defines the dataset, approach to predictive modeling, and performance evaluation. Furthermore, the fourth section details feature and feature set assessment methods. The fifth section outlines the collected outcomes critically examined in the last part of the paper.

Corresponding author:
Martin Fridrich, BUT, Faculty of Business and Management, Kolejní 2906/4, 612 00 Brno, Czech Republic
Email: fridrichmartin@yahoo.com

Literature Review

To outline the relevant research endeavours, we concentrate on peer-reviewed articles from the Web of Science and Scopus databases, dealing with user/customer churn in the e-commerce environment (see Table 1). The studies determine customer churn in non-contractual settings analytically, and neither probability models (Fader et al., 2005; Netzer et al., 2008) nor cost-benefit frameworks (Glady et al., 2009; Clemente-Císcar et al., 2014) are employed. Gordini and Veglio (2017), Li and Li (2019), Chou and Chuang (2018), and Llave Montinel and López (2020) use the subsequent periods without financial transactions to ascertain customer churn events. Rachid et al. (2018) extend this notion by combining defection levels with changes in transactional behavior. However, Abbasi et al. (2015), Lee et al. (2018), Rothmeier et al. (2020), and Berger and Kompan (2019) recognize the churn event in the e-commerce domain as dropping out of the buying process or defecting across web or game sessions.

Gordini and Veglio (2017) and Li and Li (2019) deem transactional features (e.g., last transaction date, number of financial transactions, total revenue) the most relevant and completely omit the domain-specific ones. By contrast, Rachid et al. (2018), Berger and Kompan (2019), Rothmeier et al. (2020), and Perisic and Pahor (2021) show that the combination of transactional and behavior/usage (e.g., number of sessions, session length, conversion rate) can significantly improve the predictive performance of the model. Other appealing characteristics, such as perceptional features (Almuqren et al., 2021), decision stages (Abbasi et al., 2015), or geospatial patterns (Llave Montinel and López, 2020), are included. However, the evidence of their importance may be anecdotal.

The building blocks of the modeling pipeline involve data processing, sampling, dimensionality reduction, or modeling and evaluation, depending on the researcher's objectives. Nonetheless, most articles concentrate on the last step, favouring a train–test split experimental design and confusion matrix-based performance metrics (Yu et al., 2011; Kim and Lee, 2012; Abbasi et al., 2015; Llave Montinel and López, 2020, Chou and Chuang, 2018; Almuqren et al., 2021). The prevalent classification algorithms include logistic regression (LR), support vector machine (SVM), artificial neural networks (ANN), and meta-learners.

Table 1. Selected literature on user/customer churn prediction in e-commerce.

Research Article	Experiment	Modeling	Performance Metrics
Abbasi et al. (2015)	train–test splits	SVM, Bayesian nets	ACC, PRE, REC, F1
Almuqren et al. (2021)	train–test splits	ANN	PRE, REC, F1
Berger and Kompan (2019)	cross-validation	Birch, K-means, SVM	ACC, PRE, REC, AUC
Chou and Chuang (2018)	train–test splits	GAM, Bagging, Boosting	AUC
Gordini and Veglio (2016)	cross-validation, train–test splits	LR, SVM, ANN	ACC, AUC, LIFT
Kim and Lee (2012)	train–test splits	LR, ANN, Bagging	REC, AUC, LIFT
Li and Li (2019)	train–validation–test splits	LR, Gradient boosting	ACC, PRE, REC
Llave Montinel and López (2020)	train–test splits	GAM, Spatial probit	1-ACC, AUC
Perisic and Pahor (2021)	cross-validation	LR, Bagging	AUC
Rachid et al. (2018)	cross-validation	Decision tree, ANN, Bagging	ACC, PRE, REC, F1
Rothmeier et al. (2020)	cross-validation	LR, Decision tree, SVM, ANN, Bagging, Boosting, Others	AUC
Yu et al. (2011)	train–test splits	Decision tree, SVM, ANN	ACC, PRE, REC, LIFT

Source: Authors

User Churn Model

To propose a viable user churn model, we adopt a practical suggestion by Tamaddoni Jahromi et al. (2014) and use the accessible, fundamental attributes while maximizing the model's predictive power. We form the model around user-item interactions, data generally available to any e-commerce retail store.

The churn event (dependent variable) is characterized as interaction/no interaction with the e-commerce website during the subsequent month; thus, the explained variable is binary. This interpretation allows us to reflect on online behavior over a considerable period. Other works define churn ordinarily as transactional (Gordini and Veglio, 2017; Li and Li, 2019) or employing real-time interactions (Abbasi et al., 2015; Berger and Kompan, 2019).

The user model (independent variables) consists of six sets of attributes: recency, frequency, monetary, category and item, date–time, and other characteristics (see Table 2). The first three sets cover established behavioural perspectives. Hughes (2012) describes them as Recency—the time since the last transaction, where a short span indicates a high probability of returning customer; Frequency—the number of transactions within a period, where high frequency corresponds to increased loyalty; and Monetary—the total revenue within a period, where high spending marks high valued customers. We extend the outlined notions to the online environment with session- and interaction-level features. The scarcity of purchasing data also drives this shift. Matching dimensions are present across all reviewed articles. Gordini and Veglio (2017) and Li and Li (2019) acknowledge that transactional characteristics are the most influential. On the other hand, Berger and Kompan (2019) and Rachid et al. (2018) demonstrate that the amalgam of transactional and e-commerce behavior/usage leads to notable gains in classification performance.

Category & item set aims to capture a user's preference. Interest in a vast range of products may be a sign of a loyal customer (Mozer et al., 2000); issues in a particular category, on the other hand, may lead to customer defection (Buckinx and Dirk, 2005). While research endeavors support categorical behavior (Gordini and Veglio, 2017), its impact on user/customer churn in the online context seems understudied.

Date & time attributes reflect the prospect of a different user experience level during the day (Buckinx and Dirk, 2005), e.g., peaking URL requests during noon may lead to an inadequate response time and worsen the experience. Berger and Kompan (2019) consider comparable attributes for real-time sessions; however, the authors do not evaluate the variable's importance directly.

Others include time-to-event features and average session length. We expect a long session or a short time between interactions to indicate a more engaged user. Again, Berger and Kompan (2019) introduce a similar perspective.

Table 2. User Churn model.

Set	Attribute	Description	Variable	Data Type
Recency	session recency	time difference from the last user session and current (split) date [days]	ses_rec	float
	average session recency	average period between sessions [days]	ses_rec_avg	float
	standard deviation in session recency	standard deviation in time between sessions [days]	ses_rec_sd	float
	cv session recency	ratio of standard deviation in time to session to average time to session (coefficient of variation) [%]	ses_rec_cv	float
	user maturity	difference between the start of the first user session and current (split) date [days]	user_rec	float
Frequency	session frequency	session count [n]	ses_n	int
	relative session frequency	ratio of session frequency to account maturity [session/a day]	ses_n_r	float
	user–app interaction frequency	user–application interaction (view/add-to-cart/buy clicks) count [n]	int_n	int
	relative user–app interaction frequency	ratio of user–app interaction frequency to session frequency [int/session]	int_n_r	float
	transactional frequency	transaction count [n]	tran_n	int
	relative transactional frequency	ratio of transactional frequency to session frequency (individual conversion rate) [transaction/session]	tran_n_r	float
Monetary	transactional revenue	total revenue [USD]	rev_sum	float
	relative transactional revenue	total revenue to session frequency [USD/session]	rev_sum_r	float
	above average transactional revenue	proportion of sessions with above-average spending [%]	major_spend_r	float
Category & item	interactions across root-level categories	sum of interactions across root-level categories [n]	int_cat1_n: int_cat24_n	int
	relative user-cat interaction frequency	average no of distinct root-level categories interacted in session [n]	int_cat_n_avg	float
	relative user-item interaction frequency	average no of distinct items interacted in session [n]	int_itm_n_avg	float

Date & time	average month	average month (session start)	ses_mo_avg	float
	standard deviation in months	standard deviation in months	ses_mo_sd	float
	average day-hour	average hour of a day (session start)	ses_hr_avg	float
	standard deviation in day-hours	standard deviation in hours of a day	ses_hr_sd	float
	weekend proportion	weekend sessions proportion [%]	ses_wknd_r	float
Others	average session length	average session duration [min]	ses_len_avg	float
	time to interaction	average time between interactions within a session [mins]	time_to_int	float
	time to transaction	average time between transactional events [days]	time_to_tran	float

Source: Authors

Research Methodology and Implementation

In this section, we describe the components supporting the indirect assessment of the user churn model, namely (1) the dataset and its properties, (2) building blocks of modeling pipelines, and (3) feature importance estimation procedures. Furthermore, we accompany the work with code and data repositories to ensure transparency and reproducibility; see the Supplementary section.

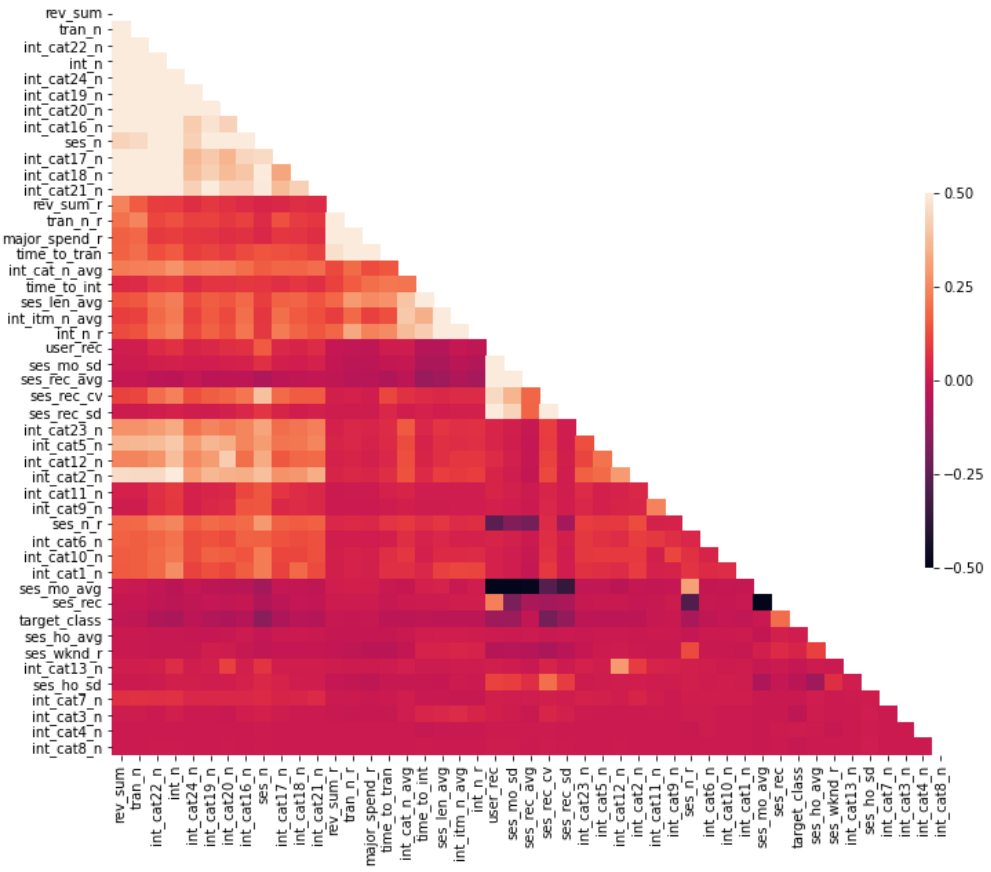

Fig. 1. Lower triangular matrix of Pearson's correlation coefficients within the user churn model.

Source: Authors

Dataset

We introduce an open e-commerce retail dataset based on the Retail Rocket Dataset. It covers the period of 2015/05/09–2015/09/17 and consists of 49,358 observations, 47 user model vectors, and the churn event. The target class distribution is imbalanced, with a churn rate of ~89%.

Further univariate exploratory analysis reveals that most attributes are asymmetric and suffer from outliers; Monetary and Category & item characteristics exhibit sparsity. The multivariate investigation reveals multicollinearity inside the user churn model; see the lower triangular matrix of Pearson's correlation coefficients in

Figure 1. The elements are sorted using agglomerative clustering with Ward linkage, which allows us to see underlying groups of associated features, e.g., the light triangle in the upper-left corner represents a coherent cluster of positively correlated features, which establishes a weak negative association with the lower-right set of features, including the target class. See the Supplementary Materials section for further details.

Modeling Pipeline

Data Processing

The transformation steps ordinarily ascertain that the basic assumptions of the downstream techniques are satisfied. Following the exploratory analysis, we concentrate on the explanatory variables. We impute missing values, omit the vectors with near-zero variance, and add a second-degree polynomial expansion to reduce the bias. Furthermore, we approach the disparity in unit measurements and irregularity in the shape of the observed probability distributions with a uniform quantile transformation.

Feature Extraction

To mitigate multicollinearity and sparsity amongst the explanatory variables, we apply the principal component analysis technique to project the processed data into orthonormal 50-dimensional space while capturing the most variability. Additional benefits may include improvements in the predictive performance and generalization and reductions in computational runtime and memory requirements (Aggarwal, 2014).

Modeling

There is a substantial amount of research devoted to the algorithm selection problem. Nevertheless, we decided to employ classification methods prevalent in the user/customer churn domain (see Table 1). For the most part, we use out-of-the-box implementations and hyperparameter settings hinged on the scikit-learn library (Pedregosa et al., 2011), with SVMs being an exception (see Table 3).

Table 3. Classification algorithms and implementation notes.

Family	Algorithm	Implementation Notes	Abbreviation
generalized linear models	regularized logistic regression	L2 penalty	LR
support vector machines	support vector machine with a linear kernel	L2 penalty, Platt's method for probability estimates	SVM-LIN
support vector machines	support vector machine with a radial basis kernel	explicit mapping with Nystroem kernel, L2 penalty, Platt's method for probability estimates	SVM-RBF
artificial neural networks	multi-layer perceptron	1 hidden layer with 100 RELU units, stochastic gradient descent, 200 epochs, L2 penalty	MLP
meta-learning (bagging)	random forest	100 decision trees, Gini criterion	RF
meta-learning (boosting)	gradient boosting machine	100 decision trees, Friedman MSE criterion	GBM

Source: Authors

Performance Benchmark

Experimental Design

To secure a reliable performance benchmark, we utilize a stratified 20-fold cross-validation scheme (see Figure 2). Firstly, we obtain training and validation data partitions. The modeling pipeline is instantiated and fitted using the training split, and its performance is assessed on both data partitions. The procedure is repeated until each split acts as a validation set; intermediate pipeline and outcomes are collected and stored.

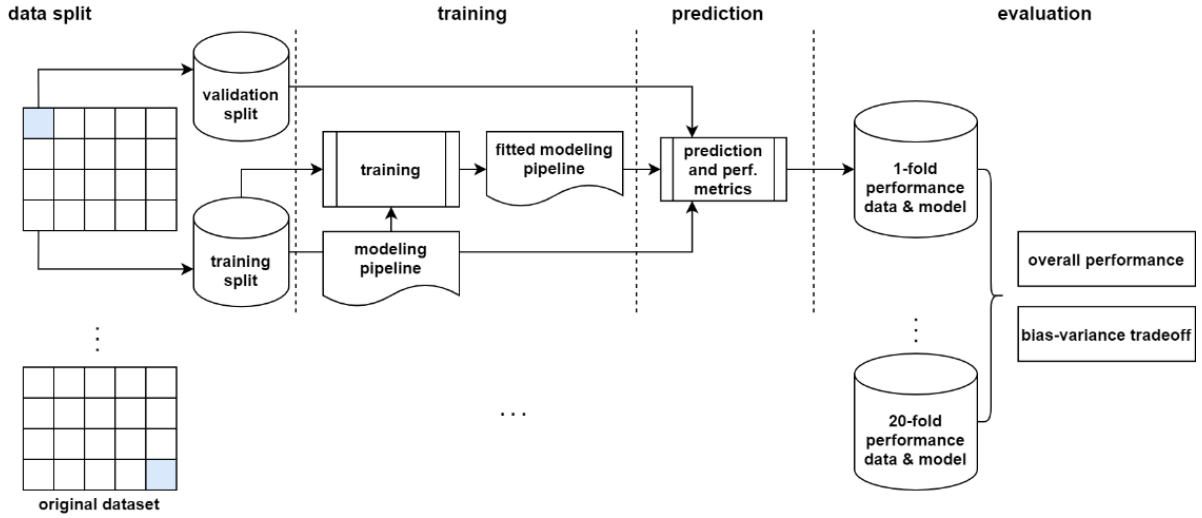

Fig. 2. Performance benchmark experimental design.
Source: Authors

Classification Metrics

We assess the classification capabilities with well-known confusion-matrix-based and subjective independent measures. We select Accuracy for its comprehensibility and reflect on the distribution imbalance in the target class with the F1 score (harmonic mean of the Precision & Recall). Furthermore, we select the threshold-independent Area under the receiver operating characteristic curve. This metric describes the probability that randomly drawn members of the retained class will produce a lower churn score than randomly drawn churners. More specifically, let us have a class membership score $s = s(x)$, as a function of user model vector x; the probability density function of corresponding scores $f_k(s)$, with a cumulative distribution $F_k(s)$, and classes $k \in \{0,1\}$; then,

$$AUC = \int_{-\infty}^{\infty} F_0(s) f_1(s) ds. \quad (1)$$

Thus, it reflects on the learner's ranking capacity. However, AUC also suffers from a few conceptual shortcomings (Hand, 2009).

Feature Importance

This article strives to assess individual feature importance and feature set importance. To address the former, we use the permutation technique Breiman (2001) proposed. It measures the variation in the performance of the modeling solution when a feature is randomly reordered. We construct the empirical probability distribution for each explanatory variable using twenty validation splits and ten repeated permutations.

We propose an original approach based on the information retrieval theory to address the latter. We treat each feature set as retrieved documents and the most influential features identified in the previous step as relevant documents, which allows us to evaluate the set importance using the following. Let us have a set of features G and a set of the n most influential features F_n; then,

$$PRE_G = \frac{|G \cap F_n|}{|G|}, \quad (2)$$

$$REC_G = \frac{|G \cap F_n|}{n}, \quad (3)$$

$$F\beta_G = (1 + \beta^2) \cdot \frac{PRE_G \cdot REC_G}{(\beta^2 \cdot PRE_G) + REC_G}, \quad (4)$$

where PRE_G indicates the proportion of the relevant features within the set; REC_G shows the ratio of the influential elements captured; $F\beta_G$ allows us to combine both perspectives with the weighted harmonic mean. We consider the ten most influential features and apply a non-preferential variant of $F\beta_G$, with $\beta = 1$. As a result, we can compose empirical probability distributions for the information retrieval measures. In addition, we introduce their referential counterparts based on 10,000 random permutations in the feature importance rank.

Results

This section centres around (1) classification performance and runtimes and (2) individual and set feature

importance. We analyse the modeling pipelines and choose the most suitable ones for the feature permutation assessment, which is the article's main focus. The analysis is supported by estimates of central tendencies, confidence bounds, and hypothesis testing with Bonferroni corrections on untreated $\alpha = 0.01$.

Performance Benchmark

We compare the prediction ability of the modeling pipelines using the mean point estimates of the classification metrics and confidence bounds for the underlying distributions in Table 4. In addition, we evaluate the difference in performance for each combination of pipelines using a paired t-test coupled with validation splits. The null hypothesis claims that the true difference in sample means equals zero; the alternative hypothesis states that the actual difference in sample means is not equal to zero, i.e., there is a statistically significant difference in classification performance between the two solutions.

Table 4. Classification performance metrics computed over the validation data partitions.

Algorithm	Training time [s]	Prediction time [s]	ACC (95% CI)	F1 (95% CI)	AUC (95% CI)
LR	6.23 (±0.52)	0.09 (±0.01)	0.8894 (±0.0010)	0.9410 (±0.0005)	0.7374 (±0.0082)
SVM-LIN	109.19 (±8.19)	0.08 (±0.01)	0.8888 (±0.0010)	0.9406 (±0.0005)	0.7315 (±0.0076)
SVM-RBF	14.65 (±1.22)	0.12 (±0.01)	0.8879 (±0.0011)	0.9401 (±0.0006)	0.7345 (±0.0079)
MLP	46.62 (±3.64)	0.10 (±0.02)	0.8762 (±0.0018)	0.9327 (±0.0010)	0.6869 (±0.0086)
RF	68.19 (±3.7)	0.22 (±0.02)	0.8865 (±0.0010)	0.9391 (±0.0005)	0.7102 (±0.0085)
GBM	103.87 (±6.01)	0.08 (±0.01)	0.8894 (±0.0010)	0.9409 (±0.0005)	0.7426 (±0.0081)

Source: Authors

The LR, SVM, and GBM algorithms are almost on par in the threshold-dependent confidence matrix-based ACC and F1, with LR being the best-performing classifier. We fail to reject the null hypothesis for the pairs and metrics, i.e., there does not seem to be enough evidence to distinguish between the solutions. The outcomes of subject-independent AUC are more diverse, with GBM achieving the highest value. We still fail to reject the null hypothesis for associations amongst LR, SVM-RBF, and GBM. The remaining solutions do not perform well; the gap is statistically significant and aligned with acceptance of the alternative hypothesis in most (RF) or all paired comparisons (MLP).

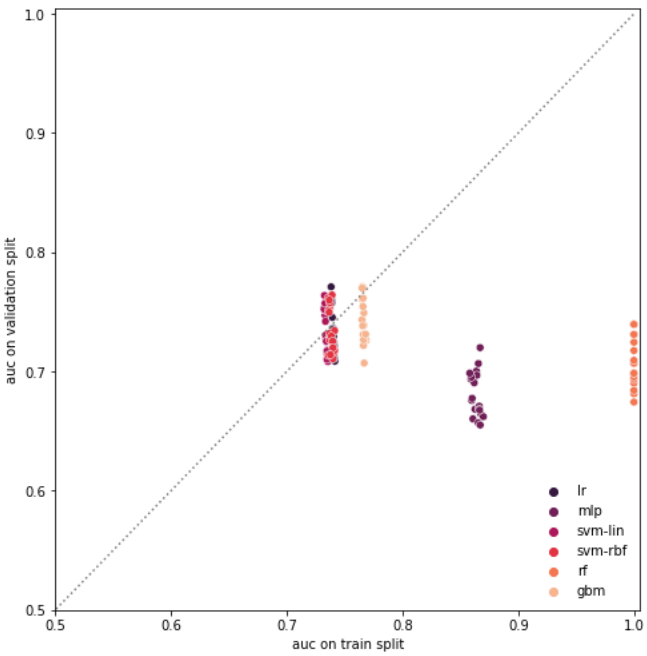

Fig. 3. Bias and variance trade-off in AUC.

Source: Authors

To understand the underwhelming performance of RF and MLP, we present the bias-variance trade-off plot in Figure 3, concentrating on AUC on both seen and unseen data. We recognize that both algorithms exhibit outstanding classification ability on training splits (low bias); however, they fail to generalize on validation splits (high variance). In other words, their out-of-the-box hyperparameter settings lead to unreasonably complex models. Nonetheless, hyperparameter optimization is beyond the scope of this article.

The disparities in computational runtime reveal additional practical insights, i.e., the training LR pipeline is faster than GBM by a factor of ~16 while yielding equivalent classification performance. In addition, we see that the Nystroem kernel mapping (Williams and Seeger, 2000) improves the training phase of the SVM solution by a factor of ~7. Thus, LR and SVM-RBF may be favoured due to their parsimony and speed.

Feature Importance

Overall, permutation feature importance is evaluated on validation data partitions using the three best-performing solutions: LR, SVM-RBF, and GBM. The most influential variables are presented in Table 5, with respective mean point estimates and confidence bounds. We also verify that the decrease in AUC is positive for all of them with the bootstrap test. Variables ses_rec, ses_rec_avg, and ses_n are indispensable for all three pipelines and are aligned with our expectations, e.g., we would suspect a user with a recent session, a low average period between sessions, and a high number of sessions to be retained.

Table 5. Individual permutation feature importance.

LR		SVM-RBF		GBM	
Variable	Decrease in AUC (95% CI)	Variable	Decrease in AUC (95% CI)	Variable	Decrease in AUC (95% CI)
ses_rec	0.0731 (±0.0019)	ses_rec	0.0791 (±0.0025)	ses_rec	0.0617 (±0.0015)
ses_rec_avg	0.0256 (±0.0012)	ses_rec_avg	0.0150 (±0.0009)	ses_n	0.0157 (±0.0009)
ses_n	0.0156 (±0.0009)	ses_n	0.0118 (±0.0008)	ses_rec_avg	0.0151 (±0.0009)
int_n	0.0097 (±0.0007)	int_n	0.0088 (±0.0008)	ses_mo_sd	0.0104 (±0.0008)
user_rec	0.0070 (±0.0006)	ses_mo_sd	0.0070 (±0.0009)	ses_rec_cv	0.0093 (±0.0008)
ses_rec_cv	0.0069 (±0.0006)	int_cat17_n	0.0059 (±0.0008)	ses_rec_sd	0.0065 (±0.0007)
ses_n_r	0.0064 (±0.001)	int_cat22_n	0.0057 (±0.0006)	ses_n_r	0.0060 (±0.0006)
ses_rec_sd	0.0021 (±0.0004)	user_rec	0.0044 (±0.0005)	ses_mo_avg	0.0056 (±0.0007)
int_cat17_n	0.0021 (±0.0004)	ses_rec_cv	0.0042 (±0.0007)	int_n	0.0054 (±0.0006)
ses_mo_avg	0.0014 (±0.0002)	int_cat20_n	0.0041 (±0.0006)	user_rec	0.0044 (±0.0006)

Source: Authors

Feature set importance examines intersections between influential feature sets constructed in the previous step and each feature group employing information retrieval measures. The results are shown in Table 6, with corresponding mean estimates and confidence intervals supported by the bootstrap tests. The Recency set exhibits outstanding qualities with a very high F1; 65–86 % of its elements are identified as relevant (PRE); modeling solutions select 33–43 % of the essential characteristics from the Recency set (REC). The Frequency group is the runner-up, with a considerable F1; 36–52% PRE and 25–32% REC in LR and GBM. Category & item and Date & time characteristics display a moderate F1; the former suffers from sparsity (low PRE) and is the essential set for SVM-RBF (high REC); the latter group shows acceptable PRE and is favoured by GBM (fair REC). The remaining feature sets appear irrelevant for all modeling pipelines; however, we can statistically confirm this only for LR.

Table 6. Feature set importance.

Set	Algorithm	PRE (95% CI)	REC (95% CI)	F1 (95% CI)
Recency	LR	0.863 (±0.017)	0.432 (±0.008)	0.576 (±0.011)
	SVM-RBF	0.651 (±0.022)	0.325 (±0.011)	0.434 (±0.015)
	GBM	0.805 (±0.022)	0.402 (±0.011)	0.537 (±0.015)
Frequency	LR	0.524 (±0.016)	0.315 (±0.010)	0.393 (±0.012)

	SVM-RBF	0.362 (±0.021)	0.217 (±0.013)	0.272 (±0.016)
	GBM	0.414 (±0.016)	0.248 (±0.010)	0.310 (±0.012)
Monetary	LR	0.000 * (±0.000)	0.000 * (±0.000)	0.000 * (±0.000)
	SVM-RBF	0.002 (±0.003)	0.001 (±0.001)	0.001 (±0.001)
	GBM	0.002 (±0.003)	0.001 (±0.001)	0.001 (±0.001)
Category & item	LR	0.074 (±0.005)	0.186 (±0.013)	0.106 (±0.007)
	SVM-RBF	0.134 (±0.006)	0.335 (±0.015)	0.192 (±0.009)
	GBM	0.051 (±0.005)	0.126 (±0.012)	0.072 (±0.007)
Date & time	LR	0.137 (±0.020)	0.069 (±0.010)	0.092 (±0.013)
	SVM-RBF	0.225 (±0.021)	0.113 (±0.010)	0.150 (±0.014)
	GBM	0.406 (±0.025)	0.203 (±0.012)	0.271 (±0.017)
Others	LR	0.000 * (±0.000)	0.000 * (±0.000)	0.000 * (±0.000)
	SVM-RBF	0.027 (±0.015)	0.008 (±0.005)	0.012 (±0.007)
	GBM	0.065 (±0.024)	0.020 (±0.007)	0.030 (±0.011)

Note: * unadjusted $p < 0.01$, for $H_0: \mu > 0, H_A: \mu \leq 0$

Source: Authors

We test the difference in mean point estimates of the information retrieval measures for feature importance ranks generated by predictive solutions and their randomly permutated counterparts to explicitly compensate for group size. The outcomes are displayed in Table 7, consisting of difference estimates and confidence bounds, and further validated with the bootstrap tests. Recency and Frequency surpass their referential sets in all aspects. Date & time also performs well when coupled with GBM. The remaining feature groups seem inferior.

Table 7. Difference in mean point estimates between the observed and referential feature set importance.

Set	Algorithm	ΔPRE (95% CI)	ΔREC (95% CI)	ΔF1 (95% CI)
Recency	LR	0.651* (±0.017)	0.325* (±0.009)	0.434* (±0.011)
	SVM-RBF	0.438* (±0.021)	0.219* (±0.011)	0.292* (±0.014)
	GBM	0.592* (±0.022)	0.296* (±0.011)	0.395* (±0.014)
Frequency	LR	0.312* (±0.016)	0.187* (±0.010)	0.234* (±0.012)
	SVM-RBF	0.150 * (±0.021)	0.090 * (±0.012)	0.112* (±0.015)
	GBM	0.201* (±0.017)	0.121* (±0.010)	0.151* (±0.012)
Monetary	LR	−0.212 (±0.005)	−0.064 (±0.001)	−0.098 (±0.002)
	SVM-RBF	−0.211 (±0.006)	−0.063 (±0.002)	−0.097 (±0.003)
	GBM	−0.211 (±0.005)	−0.063 (±0.002)	−0.097 (±0.003)
Category & item	LR	−0.139 (±0.005)	−0.347 (±0.013)	−0.198 (±0.008)
	SVM-RBF	−0.079 (±0.006)	−0.197 (±0.015)	−0.113 (±0.009)
	GBM	−0.162 (±0.005)	−0.406 (±0.011)	−0.232 (±0.007)
Date & time	LR	−0.075 (±0.020)	−0.037 (±0.010)	−0.050 (±0.013)
	SVM-RBF	0.013 (±0.021)	0.006 (±0.010)	0.009 (±0.014)
	GBM	0.194* (±0.024)	0.097* (±0.012)	0.129* (±0.016)
Others	LR	−0.213 (±0.005)	−0.064 (±0.001)	−0.098 (±0.002)
	SVM-RBF	−0.186 (±0.017)	−0.056 (±0.005)	−0.086 (±0.008)
	GBM	−0.148 (±0.024)	−0.044 (±0.007)	−0.068 (±0.011)

Note: * unadjusted $p < 0.01$, for $H_0: \Delta\mu \leq 0, H_A: \Delta\mu > 0$

Source: Authors

Discussion

Individual importance revealed the ses_rec, ses_rec_avg, and ses_n variables as the most influential. Gordini and Veglio (2017) and Li and Li (2019) rely on their transactional counterparts. Rachid et al. (2018) depend on features describing users' behavior within a purchasing process. The disparities are driven by churn perception and business context, e.g., the reported studies focus on transactional churn in explained and explanatory variables. Other culprits may include learners, evaluation procedures, or the use of data partitions.

Set perspective recognized Recency and Frequency as the most important regarding information retrieval measures and matching differences across LR, SVM-RBF, and GBM. Berger and Kompan (2019) also present notable classification performance gains when expanding the base user model with analogous attributes. The Category & item and Date & time groups displayed moderate relevance when coupled with SVM-RBF or GBM; nevertheless, they were underwhelming when adjusted for the number of elements. This problem might be alleviated with variable preselection, dense encodings, and additional feature engineering. Correspondingly, Gordini and Veglio (2017) associate transactional preferences with churn events. Monetary & Other characteristics seemed inferior. There does not appear to be general agreement on the importance of the former set; e.g., Berger and Kompan (2019) present contradictory results while introducing the Monetary features.

Our findings support some of the standard dimensions of customer analytics and expose the possible value of preference and date-time behavioural patterns. Further local comprehension of suitable modeling pipelines may assist retention management professionals in leveraging imprinted inclinations to formulate personalized value propositions and campaign schedules. The global interpretation might expose problematic product segments or unsatisfactory user experience. Other directions for future research may involve a broader spectrum of business contexts. Other feature groups can inform the user model, such as geospatial or perceptual characteristics. New insights may be supported by diverse classification learners, hyperparameter optimization, and looser pipeline selection criteria. In addition, future research endeavours might divert from predictive to causal modeling and examine the structure of the underlying decision-making process. Schiffman et al. (2012) suggest enlightening such aspiration with the external influences (company and its environment), the process itself (decision stages, experience, psychological aspects), and its outputs (purchase or usage, post-evaluation); the redirection may further inform consumer marketing and behavior theory.

The model-agnostic set evaluation procedure balances aspects of effects on predictive performance and overall quality. It also accounts for interactions amongst the explained variables. Unfortunately, the impact on performance is assessed indirectly. Other shortcomings include the arbitrary size of the most influential group or the omission of its inner rankings. Thus, we suggest further expansion with comprehensive sensitivity analysis and rank-aware metrics.

Conclusions

Over the last twenty years, technological innovations and the transition towards user-centric thinking have fueled the rise of the retail e-commerce sector. It has become imperative to maintain a healthy user/customer base through retention management. Churn prediction informs both short- and long-term retention pursuits and relies upon dependent and independent attributes. However, the research literature does not demonstrate agreement on such a user/customer churn model. Thus, we proposed a user churn model suitable for e-commerce retail and investigated its properties using auxiliary evaluation. We shaped the model around interactions with the website and covered various aspects of user behavior. The indirect assessment of permutation importance was carried out on unseen data employing the best performing solutions, namely, LR, SVM-RBF, and GBM.

Individual importance acknowledged the period from the last session, the average period between the sessions, and the number of sessions as the essential features. Similarly, the set perspective recognized the importance of recency and frequency characteristics concerning information retrieval metrics and matching differences across the relevant modeling pipelines. Furthermore, SVM-RBF and GBM learners supported new feature groups such as Category & Item or Date & time. The remaining sets manifested poor associations with the classification performance. The findings might also inform retention management endeavours, e.g., personalized campaigns or user experience enhancements. Other contributions include the original set evaluation procedure and the open dataset.

Future research may address the limitations of our work with broader investigation across multiple business contexts. The user model can be informed by other feature groups. In addition, we suggest interpreting the underlying associations using partial-dependence plots, Shapley values, or surrogate models. New insights may also be supported by employing diverse classification learners, hyperparameter optimization, and looser pipeline selection criteria. In addition, future scientific efforts may turn away from predictive to causal modeling and investigate the structure of the underlying decision-making process; the shift may inform consumer marketing and behavior theory.

The set evaluation procedure might be notably extended with information retrieval metrics directly linked to the

solutions' predictive performance and inner importance ranks. Furthermore, a thorough sensitivity analysis might be advisable.

Supplementary Materials:

Code: https://github.com/fridrichmrtn/user-churn-model-ecommerce-retail

Dataset: https://www.kaggle.com/fridrichmrtn/user-churn-dataset

References

Abbasi, A., Lau, R. Y. K., & Brown, D. E. (2015). Predicting behavior. *IEEE Intelligent Systems*, 30(3), 35-43. https://doi.org/10.1109/MIS.2015.19

Aggarwal, C. C. (2014). Data classification: algorithms and applications. Taylor & Francis.

Almuqren, L., Alrayes, F. S., & Cristea, A. I. (2021). An Empirical Study on Customer Churn Behaviours Prediction Using Arabic Twitter Mining Approach. *Future Internet*, 13(7). https://doi.org/10.3390/fi13070175

Ascarza, E., Neslin, S., Netzer, O., Anderson, Z., Fader, P., Gupta, S., Hardie, B., Lemmens, A., Libai, B., Neal, D., Provost, F., & Schrift, R. (2018). In Pursuit of Enhanced Customer Retention Management: Review, Key Issues, and Future Directions. *Customer Needs and Solutions*, 5(1), 65-81. https://doi.org/10.1007/s40547-017-0080-0

Berger, P., & Kompan, M. (2019). User Modeling for Churn Prediction in E-Commerce. *IEEE Intelligent Systems*, 34(2), 44-52. https://doi.org/10.1109/MIS.2019.2895788

Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/A:1010933404324

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User Modeling and User-Adapted Interaction*, 6(2-3), 87-129. https://doi.org/10.1007/BF00143964

Buckinx, W., & Van den Poel, D. (2005). Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual FMCG retail setting. *European Journal of Operational Research*, 164(1), 252-268. https://doi.org/10.1016/j.ejor.2003.12.010

Clemente-Císcar, M., San Matías, S., & Giner-Bosch, V. (2014). A methodology based on profitability criteria for defining the partial defection of customers in non-contractual settings. *European Journal of Operational Research*, 239(1), 276-285. https://doi.org/10.1016/j.ejor.2014.04.029

Fader, P. S., Hardie, B. G. S., & Lee, K. L. (2005). "Counting your customers" the easy way: an alternative to the Pareto/NBD Model. *Marketing Science*, 24(2), 275. https://doi.org/10.1287/mksc.1040.0098

Glady, N., Baesens, B., & Croux, C. (2009). Modeling churn using customer lifetime value. *European Journal of Operational Research*, 197(1), 402-411. https://doi.org/10.1016/j.ejor.2008.06.027

Gordini, N., & 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, 100-107. https://doi.org/10.1016/j.indmarman.2016.08.003

Gronwald, K. D. (2017). Integrated Business Information Systems: A Holistic View of the Linked Business Process Chain ERP-SCM-CRM-BI-Big Data: A Holistic View of the Linked Business Process Chain ERP-SCM-CRM-BI-Big Data. Springer Berlin Heidelberg. https://books.google.cz/books?id=mSYmDwAAQBAJ

Hand, D. J. (2009). Measuring classifier performance: a coherent alternative to the Area under the ROC curve. *Machine Learning*, 77(1), 103-123. https://doi.org/10.1007/s10994-009-5119-5

Hughes, A. M. (2012). Strategic database marketing (4th ed). McGraw-Hill.

Chou, Y. C., & Chuang, H. H. -C. (2018). A predictive investigation of first-time customer retention in online reservation services. *Service Business*, 12(4), 685-699. https://doi.org/10.1007/s11628-018-0371-z

Kim, K., & Lee, J. (2012). Sequential manifold learning for efficient churn prediction. *Expert Systems with Applications*, 39(18), 13328-13337. https://doi.org/10.1016/j.eswa.2012.05.069

Li, X., & Li, Z. (2019). A Hybrid Prediction Model for E-Commerce Customer Churn Based on Logistic Regression and Extreme Gradient Boosting Algorithm. *Ingénierie des systèmes d information*, 24(5), 525-530. https://doi.org/10.18280/isi.240510

Llave Montiel, M. A., & López, F. (2020). Spatial models for online retail churn: Evidence from an online grocery delivery service in Madrid. *Papers in Regional Science*, 99(6), 1643-1665. https://doi.org/10.1111/pirs.12552

MacKenzie, I., Meyer, C., & Noble, S. (2013). How retailers can keep up with consumers. McKInsey & Company Insights. https://www.mckinsey.com/industries/retail/our-insights/how-retailers-can-keep-up-with-consumers

Mezghani, M., Zayani, C. A., Amous, I., & Gargouri, F. (2012). A user profile modelling using social annotations. In Proceedings of the 21st international conference companion on World Wide Web - WWW '12 Companion (p. 969-). ACM Press. https://doi.org/10.1145/2187980.2188230

Morgan, B. (2018). How Amazon Has Reorganized Around Artificial Intelligence and Machine Learning. Forbes. https://www.forbes.com/sites/blakemorgan/2018/07/16/how-amazon-has-re-organized-around-artificial-intelligence-and-machine-learning/

Mozer, M. C., Wolniewicz, R., Grimes, D. B., Johnson, E., & Kaushansky, H. (2000). Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on Neural Networks*, 11(3), 690-696. https://doi.org/10.1109/72.846740

Netzer, O., Lattin, J. M., & Srinivasan, V. (2008). A Hidden Markov Model of Customer Relationship Dynamics. *Marketing Science*, 27(2), 185-204. https://doi.org/10.1287/mksc.1070.0294

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12(85), 2825-2830.

https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Perisic, A., & Pahor, M. RFM-LIR feature framework for churn prediction in the mobile games market. *IEEE Transactions on Games*, 1-1. https://doi.org/10.1109/TG.2021.3067114

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), 2367-2383. https://doi.org/10.11591/ijece.v8i4.pp2367-2383

Rothmeier, K., Pflanzl, N., Hullmann, J. A., & Preuss, M. (2021). Prediction of Player Churn and Disengagement Based on User Activity Data of a Freemium Online Strategy Game. *IEEE Transactions on Games*, 13(1), 78-88. https://doi.org/10.1109/TG.2020.2992282

Schiffman, L. G., Kanuk, L. L., & Hansen, H. (2012). Consumer Behaviour: A European Outlook (2nd). Pearson Financial Times/Prentice Hall.

Tamaddoni Jahromi, A., Stakhovych, S., & Ewing, M. (2014). Managing B2B customer churn, retention and profitability. *Industrial Marketing Management*, 43(7), 1258-1268. https://doi.org/10.1016/j.indmarman.2014.06.016

Terdiman, D. (2018). How AI is helping Amazon become a trillion-dollar company. Fast company. https://www.fastcompany.com/90246028/how-ai-is-helping-amazon-become-a-trillion-dollar-company

Williams, C., & Seeger, M. (2001). Using the nyström method to speed up kernel machines. In Advances in Neural Information Processing Systems 13 (pp. 682-688). MIT Press. https://infoscience.epfl.ch/record/161322?ln=en

Yu, X., Guo, S., Guo, J., & Huang, X. (2011). An extended support vector machine forecasting framework for customer churn in e-commerce. *Expert Systems with Applications*, 38(3), 1425-1430. https://doi.org/10.1016/j.eswa.2010.07.049

Retailrocket recommender system dataset: Ecommerce data: web events, item properties (with texts), category tree. (2017). Kaggle. Retrieved June 7, 2021, from https://www.kaggle.com/retailrocket/ecommerce-dataset/metadata