이커머스 세분화된 행동 유형을 반영한 딥러닝 기반 고객 이탈 예측 모델

오병수[01], 김영재[1]
[1]서강대학교 AI・SW대학원
ohbyungsu@hyudai-autoever.com, youkim@sogang.ac.kr

Deep Learning-Based Customer Churn Prediction Model Incorporating Segmented Customer Behavior in E-Commerce

Byungsu Oh[01], Youngjae Kim[1]
[1]Graduate School Of AI・SW, Sogang University

요 약

최근 커머스 시장이 오프라인에서 온라인으로 전환되면서 온라인 고객 확보뿐만 아니라 고객 이탈 방지의 중요성이 증가하여, 정확한 고객 이탈 예측을 통한 마케팅 효율 향상과 비용 절감이 요구되고 있다. 이 연구는 이커머스 세분화된 행동 유형을 딥러닝 기반 LSTM과 BiLSTM을 활용하여 고객 이탈 예측 모델을 제안한다. 본 연구에서 제안하는 예측 모델은 사용자 행동 로그 데이터(View, Cart, Purchase)를 기반으로 이탈을 단일 기준이 아닌 행동 유형별로 차별화하여 정의하고 각 유형에 따라 고객 이탈 여부를 독립적으로 판별하는 방식을 적용하였다. 또한 LRFM, 고객 행동, 타임스탬프 등의 정보를 자질 집합으로 구성해 정교한 모델 학습이 가능하도록 설계하였다. 이후 공정한 평가를 위해 모든 실험은 동일한 TPU 환경(메모리 335GB)에서 수행되었다. LSTM과 BiLSTM 모델을 학습한 결과, View 행동에 대해 Accuracy는 각각 91.07% F1-Score는 ROC-AUC는 96.17% 및 3-layer LSTM 모델이 가장 우수한 성능을 보였다. 반면, Cart 행동에서는 Accuracy 89.24%, F1-Score 83.62%, ROC-AUC 94.05%를 달성한 2-layer BiLSTM 모델이 Purchase 행동에서는 Accuracy 91.97%, F1-Score 93.95%, ROC-AUC 97.84%를 달성한 1-layer BiLSTM 모델이 최적의 성능을 나타냈다. 이를 통해 행동 유형별로 최적화된 LSTM 및 BiLSTM 모델이 고객 이탈 예측 성능 향상에 기여함을 확인하였다.

1. 서론

고객 이탈은 제품이나 서비스를 이용하던 고객이 더 이상 거래를 하지 않거나 경쟁사로 이동하는 현상을 의미한다. 최근 커머스 시장은 오프라인에서 온라인으로 전환되고 있으며 통계청에 따르면 2023년 온라인 유통 비중은 50.5%로 오프라인을 넘어섰다[1]. 이에 따라 온라인 고객 확보뿐만 아니라 고객 이탈 방지의 중요성도 커지고 있다. 고객 이탈은 매출 감소로 이어지며 신규 고객 확보 비용은 기존 고객 유지 비용의 최소 5배에 달한다[2]. 또한 기존 고객은 더 높은 구매 빈도와 금액을 보여 반복 구매를 통한 안정적인 매출을 만든다. 이에 기업은 고객 관리 시스템(CRM)을 활용해 적립, 할인 등 다양한 마케팅을 통해 이탈을 방지하고자 하며 이를 위해서는 정확도 높은 고객 이탈 예측 모델의 구축이 필수적이다

고객 이탈에 대한 연구는 이커머스와 온라인 게임 등 다양한 분야에서 연구가 활발하게 진행되어 왔으며, SVM (Support Vector Machine), k-NN (k-Nearest Neighbor), 랜덤포레스트(Random Forest)와 같은 기계학습 기법을 주로 활용했다. 하지만 대부분의 연구는 사용자의 세분화된 행동 유형을 구분하지 않고 단일 행동 기반으로 이탈 여부를 판단하였다는 한계가 있다[3].

이커머스에서는 사용자가 단순히 상품을 조회하거나 장바구니에 담는 등 다양한 구매 여정 단계를 거치므로 단일 행동만을 기준으로 이탈을 정의할 경우 실제 이탈 가능성을 정확하게 반영하기 어렵다.

본 논문에서는 이커머스 고객의 세분화된 행동 유형별로 이탈 시점을 차별화하여 정의하고, 이를 반영한 딥러닝 기반 예측 모델을 제안함으로써 행동 흐름의 다양성을 고려한 정교한 이탈 예측을 가능하게 한다.

이커머스 고객 이탈 예측을 위해 이탈을 단일 기준이 아닌 고객 행동 유형별로 차별화하여 정의하였다. 이후 각 행동(View, Cart, Purchase)에 적합한 cutoff 기준을 설정한 후 이를 기반으로 LSTM 및 BiLSTM 모델을 활용해 이탈 여부를 예측하였다. 실험 결과 View 행동에서는 2-layers LSTM 모델이 Cart 행동에서는 2-layer BiLSTM 모델, Purchase에서는 1-layer BiLSTM이 가장 우수한 성능을 보였다. 이를 통해 행동 유형에 따라 구조를 행동 유형별로 최적화된 LSTM 및 BiLSTM이 고객 이탈 예측 성능 향상에 효과적으로 기여함을 확인했다.

2. 배경지식

2-1. 관련연구

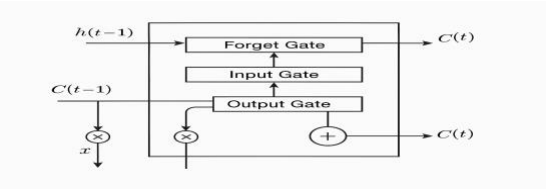
현재 이커머스 고객 이탈에 대한 연구는 활발히 진행되고 있다[4]. 하지만 이와 같은 연구는 대부분 고객의 행동을 LRFM(Length, Recency, Frequency, Monetary) 기준으로 이탈 기간을 정의하여 예측 모델을 연구하거나 혹은 데이터에 따라 이탈 기간을 정의하여 연구하였다.

최근 논문에서는 6개월(180일) 동안 구매 활동이 없으면 이탈한 것으로 정의하였고 랜덤포레스트와 AdaBoost 등 머신러닝을 이용해 예측하였다[4]. 연구를 통해 3개월 이내 구매한 고객은 유지 가능성이

높지만, 6개월 이상 미거래 고객은 이탈 가능성이 85%로 6개월 이상 구매하지 않은 고객의 이탈 가능성이 급격하게 증가된 것으로 확인하였다.[5]

2-2. LSTM(Long Short-Term Memory)

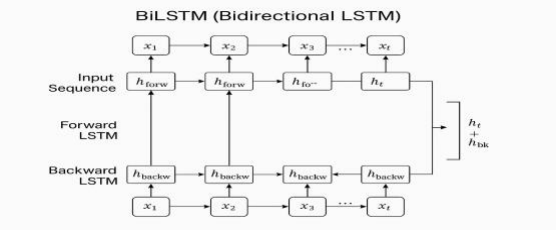
LSTM은 순서가 중요한 RNN의 한 종류로 Input gate, Output gate, Forget Gate를 통해 정보를 선택적으로 저장, 삭제, 출력하여 중요한 정보를 장기적으로 기억하고 시간에 따른 정보 흐름을 제어함으로써 시계열 데이터를 효과적으로 모델링 하는 특징이 있다.



<LSTM(Long Short-Term Memory)>

2-3. BiLSTM(Bidirectional LSTM)

BiLSTM은 LSTM을 양방향으로 확장한 구조인 RNN의 한 종류입니다. 두 개의 LSTM이 각각 정방향(Forward)과 역방향(Backward)으로 작동하며, 그 결과를 연결(concatenate)하거나 평균내어 최종 출력을 구성한다. LSTM은 과거 정보만 반영하는 반면, BiLSTM은 과거와 미래 정보를 모두 활용하여 예측 정확도를 향상시킨다.



<BiLSTM(Bidirectional LSTM)>

3. 연구제안

3-1. 연구방법

이커머스는 제품군이 다양하고 사용자 행동이 복잡하기 때문에 행동에 따라 뷰(3일 이상 행동 없으면 이탈), 장바구니 담기(7일 이내 장바구니 혹은 구매가 없으면 이탈), 구매(10일 이내 구매가 없으면 이탈)로 정의한다. 또한 반복적인 고객 활동을 기반으로 한 자질(feature) 추출이 중요하다. 특히 고객의 구매, 뷰, 장바구니 담기 등 다양한 행동 패턴을 정량화하여 자질집합으로 구성하는 것이 효과적이다. LRFM과 고객 행동, 타임스탬프 등을 자질집합으로 활용해 정교한 학습이 가능해진다. [표1]은 이커머스에서 유의미한 자질 집합을 총 25개 선정하였다.

Dimension	Name	Description
LRFM	Length	첫 접속일부터 경과일수
	Recency	마지막구매이후 경과일수
	Frequency	구매횟수
	Monetary	총 구매금액
Customer Behavior	total number of event	발생한 이벤트 총 횟수
	total number of session	발생한 세션의 총 횟수
	average event per session	접속길이
	daylength	세션길이
	sslength	세션길이
	iseveryweek	연속이용주차
	weekly change in recent session duration	최근 일주일간 세션길이 변화
	weekly change in event total count	최근 일주일간 이벤트 총 횟수
Customer Interaction	in cart	장바구니 담은 횟수
	session to purchase ratio	세션 중 구매 비율
	add to cart ratio	세션 중 카트담기 비율
	average number of sessions after purchase	구매 이후 평균 세션 수
Customer Preference	num of category	열람한 카테고리 개수
	num of product	열람한 상품 개수
	price	열람한 제품의 평균가격
Timestamp	date	이용날짜
	weekend	주말여부
	morning	시간대(06:00~12:00)
	afternoon	시간대(12:00~18:00)
	night	시간대(18:00~06:00)
	is_7daysago	마지막 7 일간 접속 여부

<표1 : 자질집합 정의>

3-2. 데이터 셋

본 연구에 활용한 데이터 셋은 Kaggle에 eCommerce behavior data from multi category store로 2019년 10월 ~ 2020년 4월까지 7개월간 약 4억건의 데이터이며 각 행은 이벤트와 제품 및 사용자와 관련이 있고 각 이벤트는 제품과 사용자 간의 다대다 관계를 갖는다.

3-3. 이탈정의

본 연구에서는 이커머스 멀티 카테고리 특성상 구매 전환 주기와 고객 행동 패턴을 고려해 행동 유형에 따라 이탈 여부를 독립적으로 판단하는 방식을 적용하였다. 각 행동 유형별로 user_id와 event_type 조합 단위로 이탈 여부를 판별하고, 고객 관심 유지 가능 기간과 구매 전환 주기를 반영하여 상이한 cutoff 기준을 설정하여 적절한 예측 성능과 마케팅 타이밍 확보가 가능하도록 정의했다.

3-4. 데이터 전처리

데이터 셋을 바탕으로 관측기간(OP)과 예측기간(CP)을 나누었다. OP 기간은 2019.11.15 ~ 2019.12.15 이고 CP기간은 2019.12.16 ~ 2020.1.14일로 정했다. 추출한 결과는 다음과 같다.

event_type	cohort	total	churned	not_churned	churn_rate (%)
cart	CP	586,559	202,075	384,484	34.45%
cart	OP	798,577	255,149	543,428	31.95%
purchase	CP	330,490	223,656	106,834	67.67%
purchase	OP	387,094	255,408	131,686	65.98%
view	CP	2,675,125	1,103,189	1,571,936	41.24%
view	OP	2,625,659	828,206	1,797,453	31.54%

<표2 : 이탈 정의에 따른 이탈 결과>

이를 바탕으로 [표1]의 자질집합과 병합하여 LSTM과 BiLSTM의 적용했다.

3-5. 연구 설정

LSTM과 BiLSTM은 파라미터를 설정 할 수 있어 주요 파라미터 값은 [표3과]같다. View는 데이터가 많아 배

치사이즈와 ephochs를 늘렸고 배치사이즈도 1024로 설정했다. Unit도 기본적으로 많이 설정하는 값을 사용해 학습했다. Cart나 Purchase는 View에 비해 상대적으로 데이터가 적어 기본 설정값을 사용하였다.

파라미터	값
EPOCHS	5 (Cart/Purchase), 10 (View)
BATCH_SIZE	512 (Cart/Purchase), 1024 (View)
LEARNING_RATE	0.001 (Cart/Purchase), 0.0005 (View)
LSTM_UNITS_1	256
LSTM_UNITS_2	128
LSTM_UNITS_3	64
DROPOUT_RATE	0.25 (Cart/Purchase), 0.3 (View)

<표3 : 파라미터 정의>

4. 연구결과

실험은 Google Colaboratory Pro에서 v2-8 TPU 환경에서 수행하였다. View LSTM과 BiLSTM 학습 결과는 [표4] 이탈 결과를 보면 알 수 있다. 비 이탈자는 정확한 것을 볼 수 있으며 이탈자 에측도 좋은 성능을 보여주고 있다. View의 가장 우수한 모델로 Accuracy 91.07%, F1-Socre 88.56%p, ROC-AUC 96.17p%가 나온 3-layers LSTM인 것을 확인 할 수 있다.

Model	Precision (비이탈자)	Recall (비이탈자)	F1 Score (비이탈자)	Precision (이탈자)	Recall (이탈자)	F1 Score (이탈자)
1-layer LSTM	0.89	0.96	0.92	0.93	0.84	0.88
2-layers LSTM	0.89	0.96	0.92	0.94	0.83	0.88
3-layers LSTM	0.89	0.96	0.93	0.94	0.84	0.89
1-layer BiLSTM	0.89	0.96	0.93	0.94	0.84	0.88
2-layers BiLSTM	0.87	0.93	0.90	0.90	0.82	0.86
3-layers BiLSTM	0.89	0.96	0.92	0.94	0.82	0.88

<표4 : view 이탈결과>

Model	Accuracy	F1 Score	ROC-AUC
1-layer LSTM	0.9067	0.8809	0.9619
2-layers LSTM	0.9083	0.8824	0.9613
3-layers LSTM	0.9107	0.8856	0.9617
1-layer BiLSTM	0.9060	0.8781	0.9577
2-layers BiLSTM	0.9102	0.8848	0.9632
3-layers BiLSTM	0.9047	0.8770	0.9576

<표5 : view 성능결과>

Cart LSTM과 BiLSTM 학습 결과는 [표6] 이탈 결과를 보면 알 수 있다. Cart도 View와 같이 비 이탈자는 정확한 것을 볼 수 있으며 이탈자 에측도 좋은 성능을 보여주고 있다. 가장 우수한 모델로 Accuracy 89.24%p, F1-Score 83.62%p, ROC-AUC 94.05p%가 나온 2-layers BiLSTM 인 것을 확인할 수 있다.

Model	Precision (비이탈자)	Recall (비이탈자)	F1 Score (비이탈자)	Precision (이탈자)	Recall (이탈자)	F1 Score (이탈자)
1-layer LSTM	0.90	0.94	0.92	0.87	0.80	0.84
2-layers LSTM	0.90	0.93	0.92	0.87	0.81	0.84
3-layers LSTM	0.90	0.94	0.92	0.88	0.79	0.83
1-layer BiLSTM	0.90	0.94	0.92	0.88	0.79	0.83
2-layers BiLSTM	0.88	0.80	0.84	0.88	0.80	0.84
3-layers BiLSTM	0.90	0.93	0.92	0.86	0.80	0.83

<표6 : cart 이탈결과>

Model	Accuracy	F1 Score	ROC-AUC
1-layer LSTM	0.8913	0.8352	0.9408
2-layers LSTM	0.8908	0.8359	0.9407
3-layers LSTM	0.8918	0.8345	0.9379
1-layer BiLSTM	0.8912	0.8337	0.9376
2-layers BiLSTM	0.8924	0.8362	0.9405
3-layers BiLSTM	0.8879	0.8310	0.9376

<표7 : Cart 성능결과>

마지막으로 Purchase LSTM과 BiLSTM 학습 결과는 [표8] 이탈 결과를 보면 알 수 있다. Purchase도 Cart와 같이 비 이탈자는 정확한 것을 볼 수 있으며 이탈자 예측도 좋은 성능을 보여주고 있다. View와 Cart의 결과와는 다르게 가장 우수한 모델로는 Accuracy 91.97%p, F1-Score 93.95%p, ROC-AUC 97.84%p가 나온 1-layers BiLSTM인 것을 확인 할 수 있다.

Model	Precision (비이탈자)	Recall (비이탈자)	F1 Score (비이탈자)	Precision (이탈자)	Recall (이탈자)	F1 Score (이탈자)
1-layer LSTM	0.86	0.88	0.87	0.94	0.93	0.94
2-layers LSTM	0.85	0.87	0.86	0.94	0.93	0.93
3-layers LSTM	0.83	0.92	0.88	0.96	0.91	0.94
1-layer BiLSTM	0.85	0.92	0.88	0.96	0.92	0.94
2-layers BiLSTM	0.85	0.88	0.87	0.94	0.93	0.93
3-layers BiLSTM	0.87	0.83	0.85	0.92	0.94	0.93

<표8: Purchase 이탈결과>

Model	Accuracy	F1 Score	ROC-AUC
1-layer LSTM	0.9145	0.9355	0.9782
2-layers LSTM	0.9101	0.9333	0.9774
3-layers LSTM	0.9155	0.9360	0.9771
1-layer BiLSTM	0.9197	0.9395	0.9784
2-layers BiLSTM	0.9121	0.9345	0.9775
3-layers BiLSTM	0.9026	0.9288	0.9734

<표9: Purchase 성능결과>

5. 결론

본 논문에서는 이커머스에서 고객 이탈을 예측하기 위해 고객 이탈을 단일 기준이 아닌 고객 행동 유형에 따라 다르게 정의하고 각 행동에 맞는 cutoff 기준을 설정하여 딥러닝 기법으로 LSTM과 BiLSTM을 사용해 이탈 여부를 판별했다. 실험결과 View에서는 3-layers LSTM이 Cart는 2-layers BiLSTM 모델이 Purchase는 1-layers BiLSTM 모델이 우수한 성능을 보였다. 이를 통해 행동 유형별로 최적화된 LSTM과 BiLSTM 모델이 고객 이탈 예측 성능 향상에 기여함을 확인하였다.

참고문헌

[1] 차대운기자, 연합뉴스, 온라인으로 더 쏠린 유통...매출 비중 '역대 최고' 50.6% ,2025

[2] F.Reichheld, "Prescription for Cutting Costs". Boston, MA, USA: Bain&Company, 2014

[3] Dr. S. Brinthakumari, Customer Churn Prediction Using Machine Learning, 2024

[4] 박수연, 고객 이용 로그와 순환신경망을 활용한 이커머스 고객 이탈 예측, 2023

[5] Berke Yilmaz, Application of Customer's Churn Prediction Models in E-commerce, 2025