온라인 쇼핑몰 구매 데이터 기반 머신러닝을 활용한 이탈율 예측 모델

Churn Prediction Using Machine Learning based on E-commerce Data

이인직*
In Jik Lee

국문초록

이 연구의 주요 목적은 고객 이탈을 정확히 예측하여 기업의 고객 유지 전략을 개선하고 장기적인 수익성을 향상시키는 데 있다.

온라인 소매몰의 3,900명 고객 구매 데이터를 기반으로 RFM(Recency, Frequency, Monetary) 모델과 다양한 머신러닝 기법을 결합하여 분석을 진행하였다. 연구 결과, Random Forest 모델이 91.11%의 재현율로 가장 우수한 예측 성능을 보였으며, 구매 금액, 연령, 고객 위치, 리뷰 평점, 구매 아이템 등이 고객 이탈 예측의 주요 변수로 확인되었다. 특히, 구매 금액이 가장 중요한 예측 변수로 나타났으며, 이는 고객의 지출 수준이 이탈 가능성과 밀접한 관련이 있음을 시사한다. 데이터 불균형 문제를 해결하기 위해 언더샘플링과 오버샘플링(SMOTE) 기법을 적용하여 모델 성능을 개선하였고, 이를 통해 소수 클래스인 이탈 고객에 대한 예측 정확도를 높일 수 있었다.

RFM 점수를 기반으로 고객을 충성 고객(38.5%), 중간 가치 고객(35.7%), 이탈 위험 고객(25.8%)으로 세분화하였다. 이러한 세분화는 이탈 위험이 높은 고객 그룹에 대해서는 특별한 프로모션이나 개인화된 서비스를 제공하여 이탈률을 감소시키는 데 도움을 줄 수 있다.

여러 머신러닝 모델을 비교한 결과, Random Forest가 정확도, 정밀도, 재현율, F1 점수 등 모든 지표에서 가장 높은 성능을 보였으며 이는 Random Forest 모델이 복잡한 고객 행동 패턴을 잘 포착하고 있음을 나타낸다. 또한, ROC 곡선 분석 결과 AUC 값이 0.99로 가장 높게 나타나, 이탈 고객과 비이탈 고객을 매우 정확하게 구분할 수 있음을 확인하였다.

이러한 연구 결과는 온라인 소매업체들이 고객 이탈을 예측하고 효과적인 고객 유지 전략을 수립하는 데 유용하게 활용될 수 있을 것이다. 특히 높은 재현율을 보이는 모델을 통해 이탈 가능성이 높은 고객을 사전에 식별하고 맞춤형 마케팅 전략을 수립할 수 있다. 이는 고객 유지율을 높이고 궁극적으로 기업의 수익성 향상에 기여할 수 있다.

향후 연구에서는 더 다양한 변수와 최신 알고리즘을 적용하여 모델의 정확도를 높이고, 실제 비즈니스 환경에서의 적용 사례를 분석하여 모델의 실효성을 검증할 필요가 있다. 이를 통해 온라인 소매업계의 데이터 기반 의사결정 역량을 더욱 강화하고, 고객 중심의 비즈니스 전략을 수립하는 데 기여할 수 있을 것으로 기대된다.

주제어: 고객 이탈 예측, 머신러닝, 이커머스, 온라인 쇼핑몰, RFM 모델, 랜덤 포레스트, 데이터 불균형, 탐색적 데이터 분석, 고객 세분화

논문접수일: 2024년 8월 1일, 게재확정일: 2025년 4월 27일
* LG전자 온라인 브랜드샵 팀(Online Brand Shop Team), good_blue@naver.com

Ⅰ. 서론

1. 연구 배경

현대 소매업계는 디지털 기술의 급속한 발전과 온라인 쇼핑의 폭발적 증가로 인해 전례 없는 변화의 시기를 맞이하고 있다. 이러한 변화는 소비자의 구매 행위에 근본적인 변화를 가져왔으며, 소매업자들은 이에 대응하여 고객의 새로운 기대와 요구를 충족시키기 위해 더욱 정교하고 맞춤화된 전략을 개발해야 하는 도전에 직면해 있다.

이러한 환경에서 소비자 구매 데이터의 분석은 비즈니스 의사 결정 과정에서 핵심적인 역할을 할 수 있다. 구매자의 연령, 성별, 구매 시점, 구매 지역, 구매 주기, 회원 여부, 결제 방법, 할인율 등 다양한 변수를 포괄적으로 분석함으로써 소매업자는 소비자의 선호와 행동을 더 깊이 이해할 수 있으며 이를 통해 소비자 행동의 미묘한 변화를 포착하고 시장 트렌드를 예측하여 지역별 또는 제품별로 차별화된 마케팅 전략을 수립할 수 있다.

더 나아가, 이러한 데이터 분석은 고객 충성도를 높이고 장기적인 고객 관계를 구축하는 데에도 핵심적인 기여를 한다. 예를 들어, 구독 여부와 결제 방법과 같은 변수의 분석을 통해 기업은 고객 충성 프로그램을 더욱 효과적으로 설계하고, 고객에게 더 매력적인 구매 조건을 제공할 수 있다. 또한, 할인율과 결제 방법의 분석을 통해 소비자의 가격 민감도와 지불 선호도를 이해함으로써, 기업은 가격 정책을 조정하고 더 공정하고 투명한 거래 조건을 마련할 수 있게 되었다.

이러한 데이터 분석은 기업의 단기적 이익 확보를 넘어, 고객 만족도를 향상시키고 장기적인 고객 신뢰를 구축하는 중요한 기반이 된다. 고객의 기대와 문제점을 더 깊이 이해하고 이에 효과적으로 대응함으로써, 기업과 소비자 모두에게 긍정적인 가치를 제공하는 선순환 구조를 만들 수 있다.

결론적으로, 이러한 다양한 요소를 종합적으로 고려한 소비자 구매 행위에 대한 심층적인 분석은 비즈니스 전략의 개발뿐만 아니라, 소비자 복지와 기업의 사회적 책임을 강화하는 데 중요한 역할을 한다. 이는 고객 경험을 개선하고, 매출을 증대시키며, 궁극적으로는 지속 가능한 비즈니스 모델을 구축하는 데 기여할 수 있다.

2. 연구 목적

본 연구의 주요 목적은 현대 소매업계에서 날로 중요해지는 소비자 구매 데이터의 복잡성과 다양성을 심층적으로 이해하고, 이를 통해 소비자 행동에 대한 포괄적인 통찰력을 제공하는 데 있다. 특히, 고객 이탈율 예측 모델 개발에 초점을 맞추고 있으며, 이는 기업의 지속가능한 성장과 경쟁력 강화에 핵심적인 요소이기 때문이다.

이탈율 예측 모델을 통해 우리는 고객 유지율을 향상시키고, 마케팅 효율성을 제고하며, 기업의 장기적 수익성을 개선할 수 있다. 예측 모델은 이탈 가능성이 높은 고객을 사전에 식별하여 적절한 retention 전략을 수립할 수 있게 하며, 고객 생애 가치를 고려한 맞춤형 마케팅 캠페인을 가능하게 한다. 또한, 이탈 원인 분석을 통해 서비스나 제품의 문제점을 파악하고 개선함으로써 전반적인 고객 경험을 향상시키는 것을 목표로 한다.

이를 위해 본 연구는 구매자의 연령, 성별, 구매 시점, 구매 지역, 구매 주기, 회원 여부, 결제 방법, 할인율 등 다양한 변수를 포함한 소비자 구매 데이터를 기술통계와 탐색적 데이터 분석을 통해 심도 있는 분석과 인사이트를 제공하고 다양한 머신러닝 기법을 활용하여 소비자의 구매 패턴과 이탈 가능성을 정확히 예측하는 모델을 개발하고자 한다.

더불어, RFM 모델을 활용한 소비자 세분화를 통해 각 고객 그룹에 맞는 맞춤형 마케팅 전략과 이탈 방지 전략을 수립하여 마케팅 효율성을 높이고 고객 만족도를 개선하는 데 크게 기여할 것이다.

궁극적으로, 본 연구는 소매업계 이해관계자들에게 실질적인 가치를 제공하고, 데이터에 기반한 의사결정을 지원하며, 고객 중심의 비즈니스 전략 수립에 기여하고자 한다. 이를 통해 기업은 시장에서의 경쟁력을 강화하고, 지속 가능한 성장을 달성할 수 있을 것이다. 소비자의 복잡한 구매 행위를 이해하고 이탈 가능성을 정확히 예측함으로써, 고객 만족도 향상과 기업의 장기적 성공을 동시에 추구하는 것이 본 연구의 궁극적인 목표이다.

Ⅱ. 문헌 리뷰

소비자 이탈 예측 모델링하기 위해서는 현재 구매 데이터를 바탕으로 먼저 이탈율을 정의하는 것이중요하다. 여러 문헌을 검토한 결과, RFM(Recency, Frequency, Monetary) 모델은 이탈 예측 및 고객 세분화하는데 널리 사용되는 방법이다.

RFM 값들은 시간에 따른 고객 행동의 변화를 포착할 수 있어서 이탈 징후를 조기에 감지하는 데 도움이 되며 3가지 다차원적인 접근을 통해 고객 행동에 대한 더 포괄적인 이해가 가능하다. 특히, 구매 데이터는 고객의 거래 과정에서 쉽게 얻을 수 있으므로 매우 실용적인 방법이다.

여러 연구에서 RFM 모델과 머신러닝 기법을 결합하여 이탈 예측의 정확도를 높이고자 연구되었으며 대부분 연구에서 XGBoost, 랜덤 포레스트 등의 앙상블 기법들이 높은 예측 성능을 보였다.

특히 모델의 예측을 상세하게 설명할 수 있어 금융, 의료 등에 많이 사용되는 SHAP(SHapley Additive exPlanations) 방법을 결합하여 이탈 요인의 중요도를 파악하는 접근법이 제안되었고(Tianpei Xu et al., 2023), 또한 행동 및 인구 통계 변수를 결합한 확장된 RFMD 모델(Thanh Ho et al., 2023)을 제안하는 연구도 진행되고 있음을 확인하였다.

최근에는 단순히 이탈 여부를 예측하는 것을 넘어 고객별 맞춤 전략 수립을 위해 세분화된 고객 그룹별로 주요 이탈 요인을 파악하는 연구들이 진행되고 있다. 전반적으로 다양한 데이터와 기법을 결합한 하이브리드 모델들이 제안되고 있으며, 예측 정확도 향상과 함께 해석 가능성을 높이는 방향으로 연구가 진행되고 있음을 알 수 있었다.

Ⅲ. 연구 방법론

1. 데이터 수집 및 전처리

연구용 데이터는 Kaggle에 있는 미국 온라인 소매몰 데이터를 활용하였다. 전체 데이터 모수는 3,900명

으로 구매자의 연령, 성별, 구매 시점, 구매 지역, 구매 주기, 회원 여부, 결제 방법, 할인 여부 등 실제 온라인몰에서 실제 사용되는 다양한 변수를 포함하고 있으므로 도출된 시사점은 실제 비즈니스 현장에서 적용할 수 있다. 결측치는 제거하였으며 이탈률 분석을 위해 불필요한 변수는 제거하였으며 범주형은 레이블 인코딩(Label Encoding)을 적용하고 수치형은 표준화하였다. 이탈율 모델링을 위해 RFM 지수 및 RFM Score 변수는 새로 생성하였다.

<표 1> 변수 정의 및 설명

No	변수명	정의	비고
1	Age	고객 연령	18~70세로 인구통계학적 변수
2	Gender	고객 성별	성별에 따른 제품 선호도와 구매 패턴 파악
3	Item Purchased	구매한 품목	Sweater / Coat / Skirt / Sunglasses / Handbag등 25개 품목
4	Purchase Amount	구매한 결제 금액	$ 20 ~ $100 범위로 미국 달러(USD)로 표시
5	Location	구매한 고객 위치	미국 52개 주
6	Season	구매한 시점 절기	(봄, 여름, 가을, 겨울) 계절성 여부 파악
7	Review Rating	고객 리뷰 점수	0~5점 사이로 구매 만족도 관련 정성적 수치
8	Subscription Status	구독 여부	Yes / No : 고객 충성도 및 반복 구매 여부 파악
9	Shipping Type	배송 옵션	일반, 특송(익일배송), 매장픽업
10	Promo Code Used	쿠폰 사용 여부	Yes / No : 프로모션 할인 적용 여부
11	Previous Purchases	이전 구매 횟수	1~50회
12	Frequency of Purchases	상품 구매 주기	주1회, 격주1회, 월1회, 분기1회, 연1회
13	Payment Method	결제 수단	계좌이체, 체크카드, 신용카드, 현금, 모바일결제, 간편결제

자료출처 : Consumer Behavior and Shopping Habits Dataset (Kaggle)

2. 이탈 예측 모델

소비자 이탈 예측 모델은 이탈 예측에 널리 사용되는 RFM(Recency, Frequency, Monetary) 모델을 활용하여 RFM Score를 신규 변수로 생성하고 머신러닝 기법을 결합하여 이탈 예측 모델을 구현하였다. 특히, 모델의 정확도를 높이기 위해 데이터 불균형을 추가 검토하였으며 전처리 작업(예: 오버샘플링, 언더샘플링, SMOTE 등)을 한 후에 모델의 정확도를 개선하였다.

마지막으로 머신러닝 대표적인 모델 8가지인 로지스틱 회귀(Logistic Regression), 결정 트리(Decision Tree), 랜덤 포레스트(Random Forest), 서포트 벡터 머신(SVM), XGBoost, K-최근접 이웃(KNN), 나이브 베이즈(Naive Bayes), 다층 퍼셉트론(MLP)을 각 모델의 성능 지표(정확도, 정밀도, 재현율, F1점수)를 교차검증하여 최적의 모델을 도출하였다.

Ⅳ. 모델링

1. 기술 통계

1.1 범주형 변수 상관 관계 분석

피어슨 상관계수는 연속형 변수에 적합하지만 범주형 변수에는 적용할 수 없으므로 범주형 변수 간의 관계를 측정하는 데 특화된 카이제곱 통계량 기반의 Cramer's V 상관관계 분석을 진행하였다.

[그림 1] 범주형(카테고리형) 변수들 간의 상관관계 분석

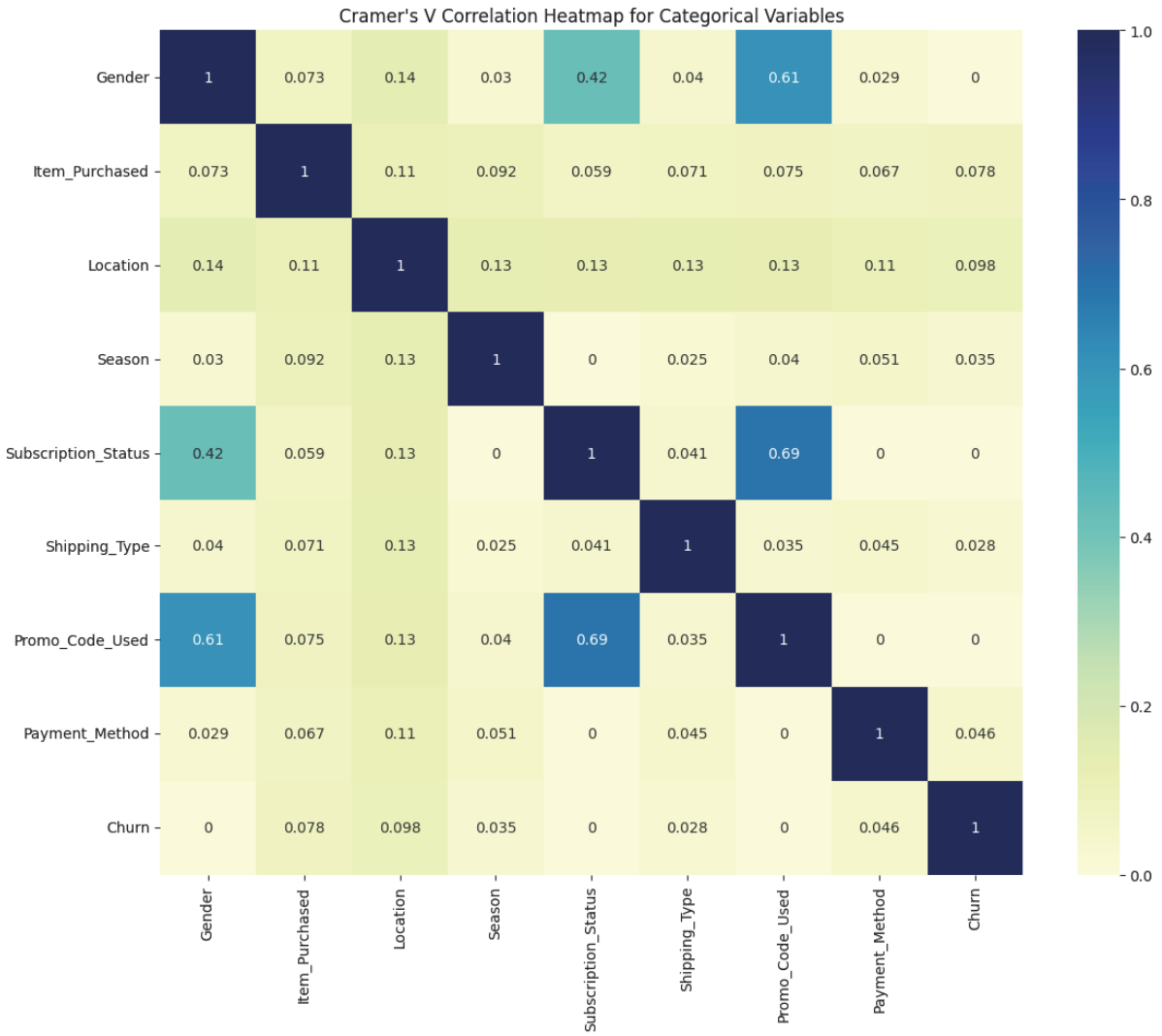


그림 1 hitmap은 Cramer's V 상관관계를 보여주는 그래프이며 주요 결과는 하기와 같다.

- Promo_Code_Used(할인 쿠폰 여부)와 Subscription_Status(구독여부)가 가장 높은 상관관계(0.69)를 보이며, 이는 구독자들은 회원 할인 등으로 프로모션 코드를 더 자주 사용할 가능성이 있다.
- Promo_Code_Used(할인 쿠폰 여부)와 Gender(성별) 사이에도 0.61로 상당히 높은 상관관계가 있다
- Gender(성별)와 Subscription_Status(구독여부) 사이에 0.42로 중간 정도의 상관관계가 있으므로 성별이 구독 상태에 영향을 미칠 수 있음을 시사한다

그 외 많은 변수가 서로 간에 낮거나 거의 없는 상관관계를 보인다(연한 노란색).

결론적으로, 대부분의 변수들이 서로 독립적인 경향을 보이며, Churn 예측에 있어 단일 변수의 강한 영향력은 보이지 않는다. 모델링 시 여러 변수의 복합적인 효과를 고려해야 할 것으로 보이며, 특히 Promo_Code_Used, Subscription_Status, Gender 간의 관계가 주목할 만하며, 이들 변수의 상호작용이 중요할 수 있다.

1.2 수치형 변수 상관 관계 분석

[그림 2] 수치형 변수 간의 피어슨 상관관계 분석

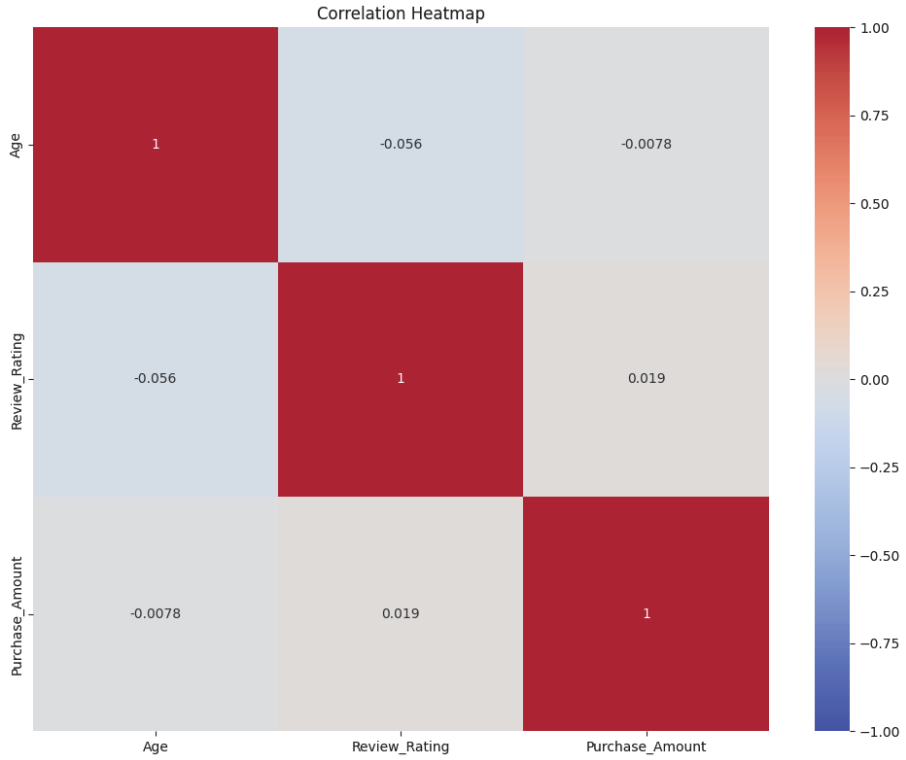


그림 2는 세 가지 변수 Age(나이), Review_Rating(리뷰 평점), Purchase_Amount(구매 금액) 사이의 상관관계를 보여준다.

Age와 Review_Rating 사이의 상관계수는 -0.056으로, 매우 약한 음의 상관관계를 보인다. 이는 나이가 증가할수록 리뷰 평점이 아주 조금 낮아지는 경향이 있음을 의미하지만 그 관계는 매우 약하다고 볼 수 있다.

Age와 Purchase_Amount 사이의 상관계수는 -0.0078로 거의 0에 가깝다. 이는 나이와 구매 금액 사이에 의미 있는 선형 관계가 없음을 나타낸다.

Review_Rating과 Purchase_Amount 사이의 상관계수는 0.019로, 매우 약한 양의 상관관계를 보인다. 리뷰 평점이 높을수록 구매 금액이 약간 증가하는 경향이 있음을 의미하지만, 그 관계 역시 매우 약하다고 볼 수 있다.

전반적으로, 세 변수 사이에는 강한 선형 상관관계가 없는 것으로 보인다. 모든 상관계수의 절대값이 0.1 미만으로, 변수들 간의 관계가 매우 약하거나 거의 없다고 해석할 수 있다.

1.3 카테고리 변수 통계 분석

실제 비즈니스 환경에서는 기본적인 통계적인 분석만으로도 중요한 인사이트 도출과 대부분의 문제점을 발굴할 수 있으므로 주기적으로 관찰해야 한다.

<표 2> 소비자 성별 분포

성별	Counts	% of Total
Male	2,652	68.0 %
Female	1,248	32.0 %

- 고객 중 남성의 비율이 여성보다 2배 이상 높다.

<표 3> 구매 주기별 빈도 분포

구매 주기	Counts	% of Total
Weekly	539	13.8 %
Bi-Weekly	1,089	27.9 %
Monthly	553	14.2 %
Quarterly	1,147	29.4 %
Annually	572	14.7 %

- 계절성 품목 구매에 따른 분기별 구매와 격주 구매가 가장 많다.

<표 4> 배송 방법별 빈도 분포

배송 방법	Counts	% of Total
Express	1,921	49.3 %
Standard	1,329	34.1 %
Store Pickup	650	16.7 %

- Express (익일 배송)가 가장 선호되는 배송 방법이다.

<표 5> 구독 여부에 따른 구매 빈도 분포

구독 여부	Counts	% of Total
Yes	1,053	27.0 %
No	2,847	73.0 %

- 대부분의 고객이 구독 서비스를 이용하지 않는다.

<표 6> 계절별 빈도 분포

계절	Counts	% of Total
Winter	971	24.9 %
Spring	999	25.6 %
Summer	955	24.5 %
Fall	975	25.0 %

- 구매는 계절에 편중되지 않고 고르게 분포되어 있다.

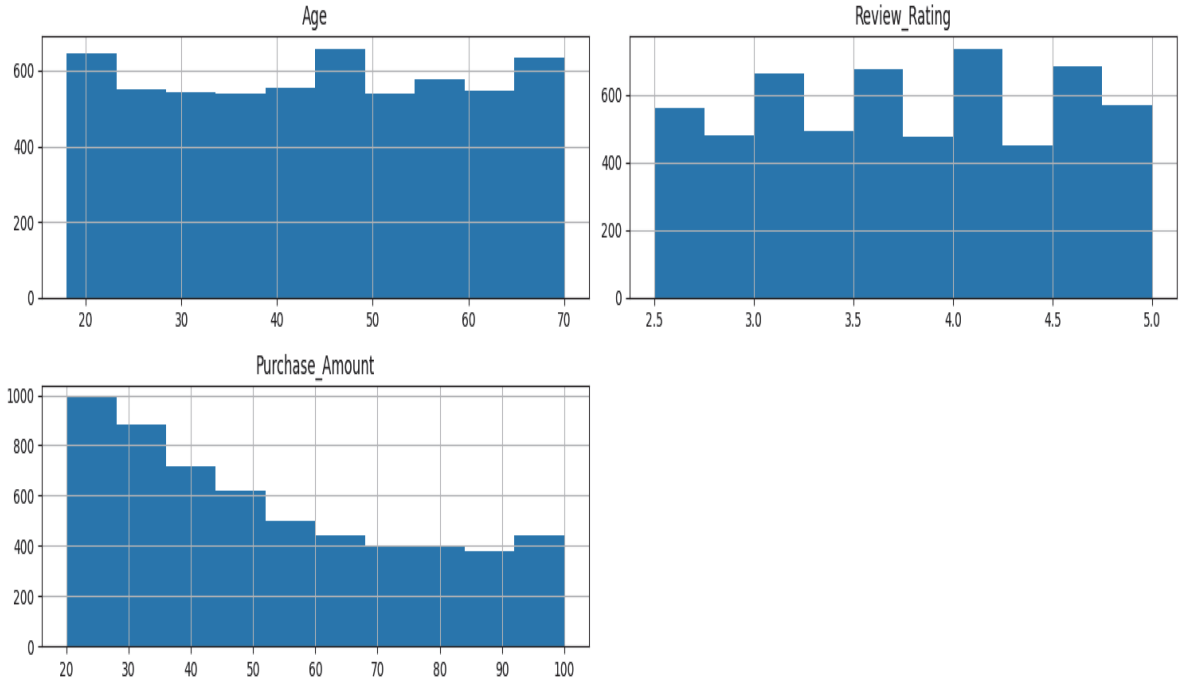
<표 7> 구매 품목별 빈도 분포

구매 품목	Counts	% of Total
Clothing	1,737	44.5 %
Footwear	599	15.4 %
Outerwear	324	8.3 %
Accessories	1,240	31.8 %

- 의류와 액세서리가 가장 인기 있는 카테고리이다.

1.4 연속형 변수 통계 분석

[그림 3] 연령, 리뷰 평점, 구매 금액별 분포



평균 나이는 약 44세이며 나이 범위는 18세부터 70세까지이다. 중앙값(50% 지점)은 44세로, 평균과 거의 일치하며 25%는 31세 이하, 75%는 57세 이하이다.

평균 구매 금액은 약 $59.76이며 구매 금액 범위는 $20에서 $100까지이다. 중앙값은 $60으로, 평균과

거의 일치하며 25%는 $39 이하, 75%는 $81 이하를 지출한다.

평균 리뷰 평점은 3.75점(5점 만점)이며 평점 범위는 2.5점에서 5점까지이다. 중앙값은 3.7점으로, 평균과 거의 일치하며 25%는 3.1점 이하, 75%는 4.4점 이상의 평점을 준다.

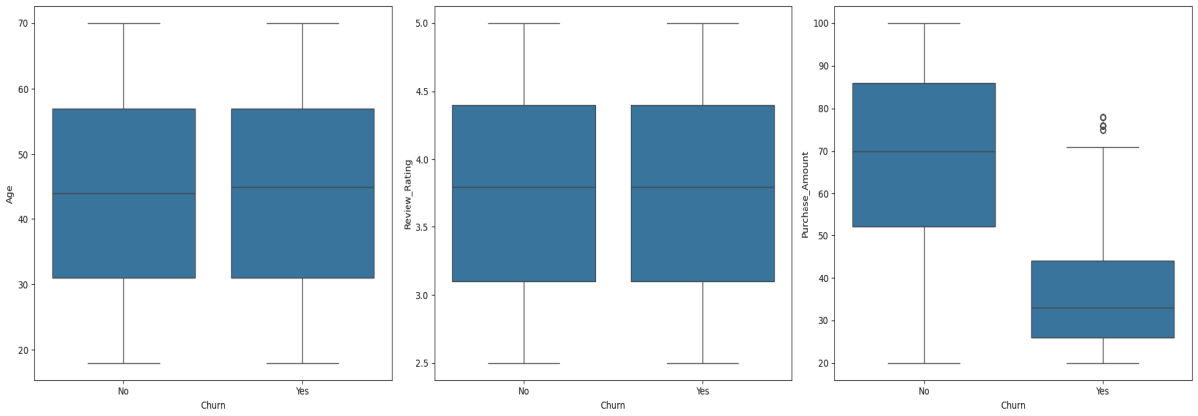
전반적인 분석 결과, 고객 연령대가 넓게 분포되어 있으며 중년층이 주요 고객층이다. 구매 금액은 비교적 고르게 분포되어 있으며, 중간 정도의 지출이 일반적이다. 리뷰 평점은 전반적으로 긍정적인 것으로 나타났다.

2. 탐색적 데이터 분석(EDA)

가장 관심이 있는 이탈율 변수를 중심으로 각 변수의 영향도를 분석하였다.

2.1 수치형 변수(연령, 리뷰평점, 구매금액)에 따른 고객 이탈 여부

[그림 4] 연령, 리뷰 평점, 구매 금액별 이탈률



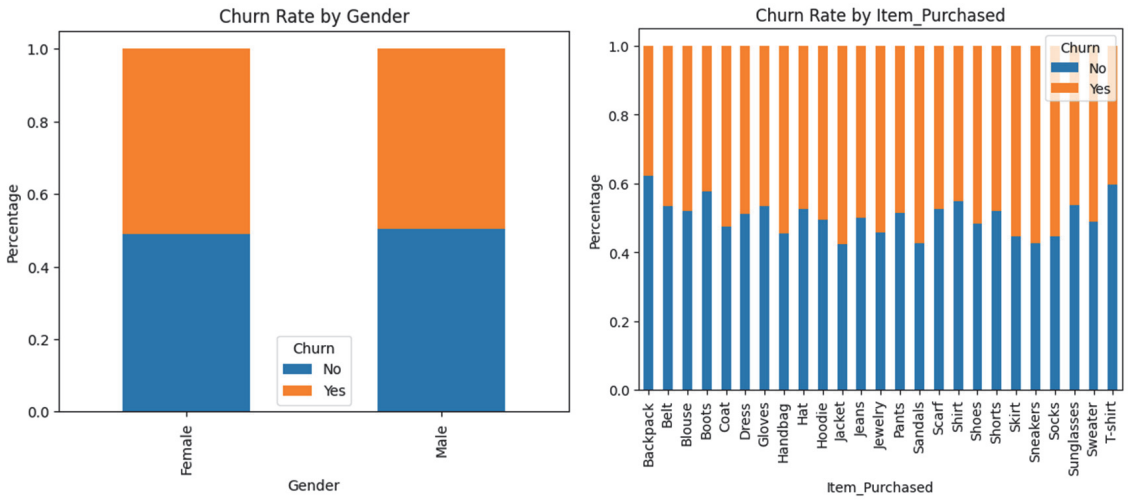
Age(나이)는 고객 이탈에 큰 영향을 미치지 않는 것으로 보이며 Review_Rating(리뷰 평점)은 대부분의 평점이 3.0에서 4.5 사이에 집중되어 있어서 고객 이탈에 큰 영향을 미치지 않는 것으로 나타났다.

Purchase_Amount(구매 금액)은 이탈 고객과 비이탈 고객 사이에 뚜렷한 차이가 보인다. 이탈 고객(Yes)의 구매 금액 중앙값이 비이탈 고객(No)보다 훨씬 낮다. 이탈 고객의 구매 금액 분포 범위가 더 좁고, 전반적으로 낮은 금액대에 집중되어 있다. 비이탈 고객의 경우 구매 금액의 분포가 더 넓고, 높은 금액대까지 분포해 있다. 따라서, 낮은 구매 금액을 보이는 고객들이 이탈할 가능성이 더 높아 보인다.

결론적으로, Purchase_Amount가 고객 이탈을 예측하는 데 가장 유용한 특성일 것으로 보인다. 나이와 리뷰 평점은 이탈 예측에 큰 영향이 없는 것으로 판단된다. 이러한 인사이트를 바탕으로 모델링 설계 시 Purchase_Amount에 더 높은 가중치를 두거나, 고객 유지 전략 수립 시 구매 금액에 초점을 맞추는 것이 효과적일 수 있다.

2.2 범주형 변수에 따른 고객 이탈 여부

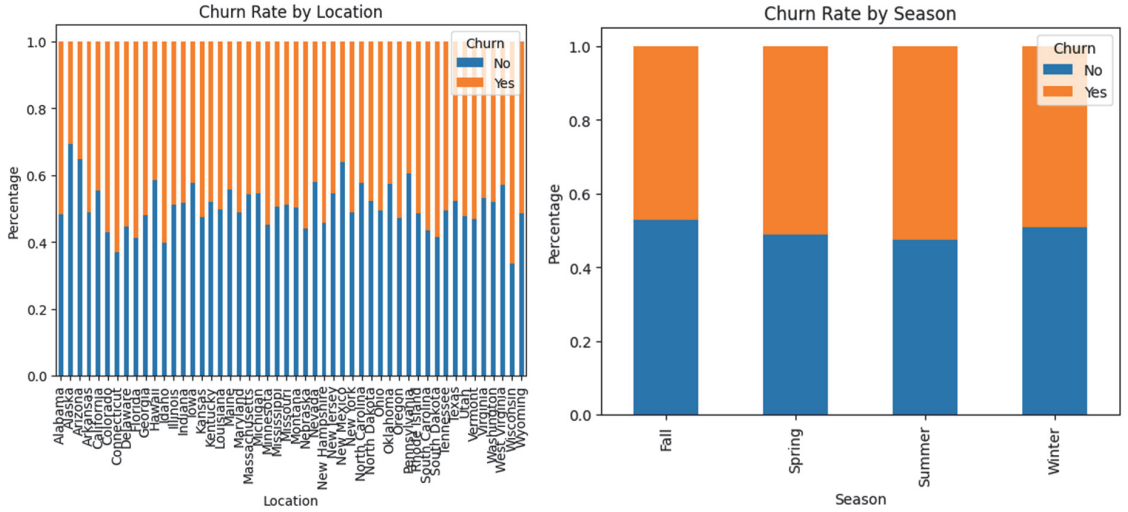
[그림 5] 성별 및 구매 품목별 이탈률



남성(Male)과 여성(Female) 모두 비슷한 이탈률을 보이므로 성별간에 고객 이탈에 큰 영향을 미치지 않는 것으로 보인다.

Item_Purchased(구매 아이템)별 이탈률에는 차이가 있다. 가장 낮은 이탈률을 보이는 아이템은 Backpack, T-shirt, Sunglasses이며, 가장 높은 이탈률을 보이는 아이템은 Jacket, Sandals, Sneakers이다. 이탈률이 높은 아이템을 구매한 고객들에게 특별한 관심을 기울이거나, 이탈률이 낮은 아이템의 판매를 촉진하는 전략을 고려할 수 있다.

[그림 6] 구매 지역과 계절별 이탈률

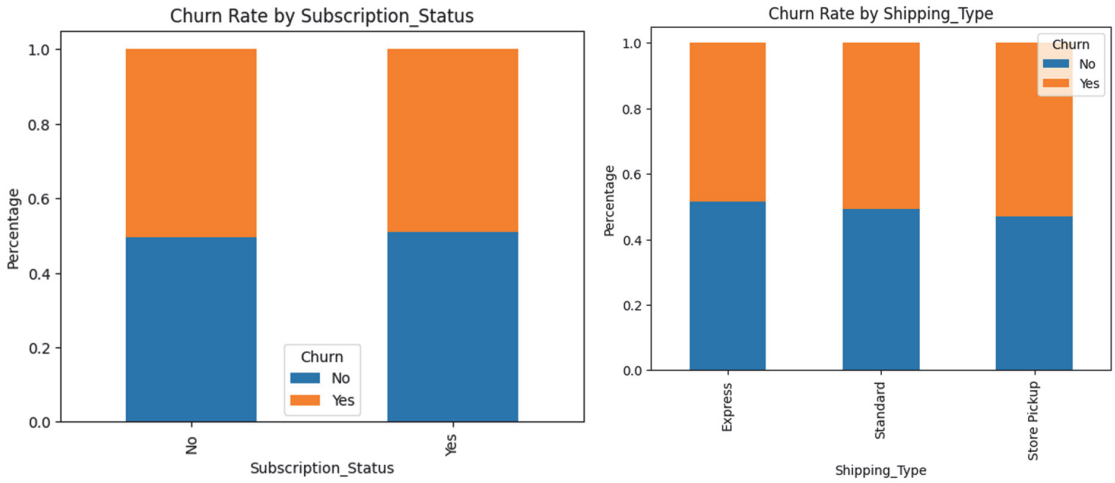


위치별로 이탈률에 상당한 차이가 있으나 대부분 이탈율은 40-60% 사이로 보인다. 일부 지역(예: Wisconsin, Connecticut, Idaho)은 더 높은 이탈율을 보이고 일부 지역(예: Alaska, Arizona, New Mexico)은 상대

적으로 낮은 이탈률을 나타낸다. 이는 지역별 특성이나 배송 등 고객 유지에 영향을 미칠 수 있음을 시사한다.

계절에 따른 이탈률 차이(최대 5% 정도)는 크지 않다. 겨울(Winter)에 이탈률이 가장 낮고 여름(Summer)에 가장 높다. 봄(Spring)과 가을(Fall)은 중간 수준의 이탈률을 보인다. 계절적 변동은 크지 않지만 여름철 고객 유지 전략을 강화할 필요가 있다.

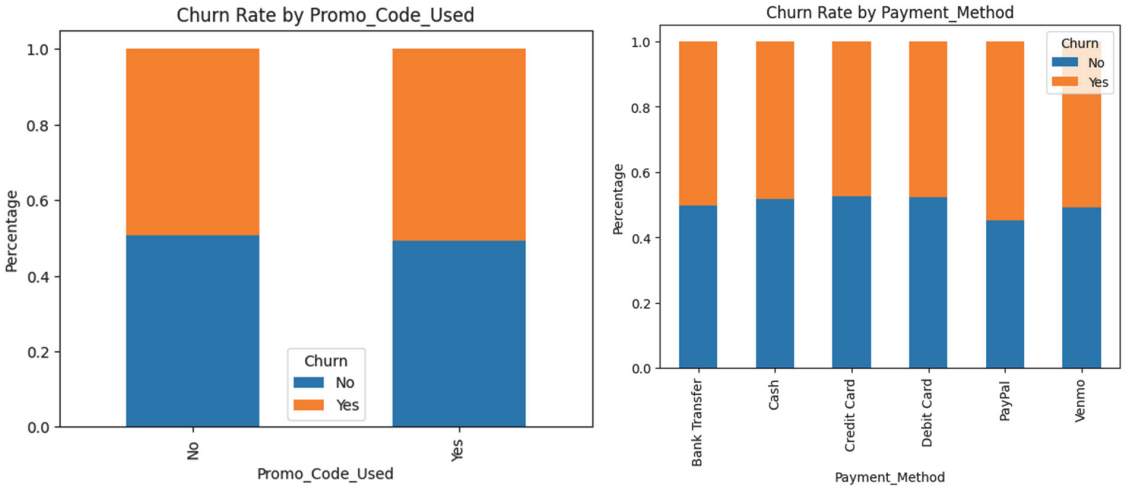
[그림 7] 구독 여부와 배송 유형별 이탈률



구독 여부와 상관없이 이탈률은 비슷한 수준을 보이므로 구독 상태가 고객 유지에 큰 영향을 미치지 않음을 시사한다.

배송 유형별로 이탈률에 차이가 있다. Store Pickup(매장 픽업)이 가장 낮은 이탈률을 보이고, Express(특송) 배송과 Standard(일반) 배송은 비슷한 수준의 이탈률을 나타낸다. 따라서, 매장 픽업 옵션이 고객 유지에 효과적이므로 매장 픽업 서비스를 강화하고 다른 배송 유형도 개선하는 것이 좋다.

[그림 8] 할인 쿠폰 사용 여부 및 결제 수단별 이탈률



할인 쿠폰 사용 여부와 관계없이 이탈률은 거의 동일하므로 고객 유지에 큰 영향을 미치지 않음을 시사한다.

결제 방법에 따라 이탈률에 차이가 있다. PayPal(간편결제)을 사용한 고객의 이탈률이 가장 높고(약 55%) Venmo(모바일결제)를 사용한 고객의 이탈률이 가장 낮다(약 45%). 다른 결제 방법들(Bank Transfer, Cash, Credit Card, Debit Card)은 비슷한 이탈률을 보인다(약 45-50%). 따라서, PayPal 결제 과정에서 문제가 있는지, 또는 PayPal 사용자들의 특성이 이탈과 연관이 있는지 추가 조사가 필요할 수 있다. 모델링 관점에서 Payment_Method 변수는 어느 정도 예측력이 있을 것으로 보인다.

3. 이탈률 예측 모델

쇼핑몰의 구매 데이터 및 인구통계학적 회원 정보를 통해 얻은 13가지 변수를 활용하여 예측 모델을 개발하겠다.

3.1 이탈 변수 생성

이탈률 변수를 만들기 위해서는 고객의 구매주기와 구매빈도를 기준으로 고객의 이탈 여부를 결정할 수 있으며 일반적으로 RFM 방법이 많이 사용된다.

1) Recency(최근성): 고객이 마지막으로 구매한 시점이 얼마나 최근인지 나타낸다. 최근에 구매한 고객일수록 다시 구매할 가능성이 높기 때문에 중요한 변수이다. 일반적으로 계산 방법은 분석 기준일(예: 오늘)과 고객의 마지막 구매일 사이의 일수를 계산한다. 일수가 적을수록 최근에 구매한 것이다. 그러나 현재 데이터에는 최근 구매 일수 변수가 없으므로, 'Frequency of Purchases(상품 구매 주기)' 변수로 대체할 수 있다.

2) Frequency(빈도): 특정 기간 동안 고객이 구매한 횟수를 나타낸다. 자주 구매하는 고객은 브랜드에 대한 충성도가 높으며, 이탈 가능성이 낮다. 일반적으로 일정 기간 동안 고객이 구매한 총 횟수를 계산한다. 본 연구에서는 Previous Purchases(이전 구매 횟수) 변수를 활용할 수 있다.

3) Monetary(금액): 특정 기간 동안 고객이 소비한 총 금액을 나타낸다. 많이 소비한 고객은 기업에 더 큰 가치를 제공하므로, 중요한 고객으로 간주된다. 일정 기간 동안 고객이 지출한 총 금액을 합산한다. 그러나 현재 데이터에는 누적 정보가 없으므로 Purchase Amount(구매한 결제 금액)변수를 활용할 수 있다.

RFM에 해당되는 3가지 변수를 통해 점수화하고 새로운 변수를 생성할 수 있다.

1) Recency(R) 점수 계산 - Frequency of Purchases
: 구매 주기가 낮을수록 점수가 낮고 이탈 가능성이 높음
Weekly : 5
Bi-Weekly : 4

Monthly : 3
Quarterly : 2
Annually : 1

2) Frequency(F) 점수 계산 - Previous Purchases
: 구매 빈도가 낮을수록 점수가 낮고 이탈 가능성이 높음
previous_purchases 40회 이상 : 5
previous_purchases 30회 이상 : 4
previous_purchases 20회 이상 : 3
previous_purchases 10회 이상 : 2
previous_purchases 10회 미만 : 1

3) Monetary(M) 점수 계산 - Purchase Amount
: 구매 금액이 낮을수록 점수가 낮고 이탈 가능성이 높음
Purchase_Amount $90 이상 : 8
Purchase_Amount $80~$89 : 7
Purchase_Amount $70~$79 : 6
Purchase_Amount $60~$69 : 5
Purchase_Amount $50~$59 : 4
Purchase_Amount $40~$49 : 3
Purchase_Amount $30~$39 : 2
Purchase_Amount $30 미만 : 1

각 구성 요소의 점수를 조합하여 RFM 최종 합계 점수를 만들고 새로운 변수를 생성하였다.
예를 들어, R=4(격주 구매), F=3(20회 이상 구매), M=4(구매금액 $50불대)인 고객의 RFM 합계 점수는 11으로 이탈 가능성이 없는 충성 고객이다.
반면에 R=2(분기별 구매), F=1(10회 이하 구매), M=3(구매금액 $40불대)인 고객의 RFM 점수는 6점으로, 이탈 가능성이 매우 높은 위험 고객, 즉 이탈 고객으로 분류할 수 있다.

합계 점수를 기준으로 크게 3가지 그룹으로 나눌 수 있다. 실제 관심 있는 그룹은 3~8점인 이탈 위험 고객이라고 할 수 있다.

- High Value Customers(충성 고객): 합산 점수 12-18
- Medium Value Customers(중간 가치 고객): 합산 점수 9-11
- Low Value Customers(이탈 위험 고객): 합산 점수 3-8

전체적인 RFM 합계 점수 분포는 하기 표와 같으며 이탈 위험 고객은 1,006명(25.8%)으로 전체의 1/4 수준이다.

<표 8> RFM 합계 점수 기반 이탈 고객 분류

Churn 구분	점수	개수	비중
충성 고객	12-18	1,503	38.5%
중간 가치 고객	9~11	1,391	35.7%
이탈 위험 고객	3-8	1,006	25.8%

3.2 이탈 예측 모델링

가장 일반적으로 많이 사용되는 랜덤 포레스트를 적용하여 예측 모델을 구축하였다. RFM Score 변수 계산에 사용된 Frequency of Purchases 및 Previous Purchases 2가지 변수는 중복으로 반영되었고 실제 모델링 활용 관점에서 불필요하므로 제외하였다.

<표 9> 분류 모델 성능 평가 지표

Classification Report

	Precision	Recall	F1-score	Support
0	0.89	0.87	0.88	596
1	0.61	**0.65**	0.63	184
accuracy			0.82	780
macro avg	0.75	0.76	0.75	780
weighted avg	0.82	0.82	0.82	780

Confusion Matrix
[[518 78]
[64 120]]

높은 recall(0.87)은 대부분 클래스 0 샘플을 정확히 식별했음을 의미한다. 반면에, 상대적으로 낮은 recall(0.65)은 클래스 1 샘플을 무작위 수준에 가깝게 분류했음을 보여준다. 전체 정확도는 82%로 보이지만, 이는 주로 다수 클래스(클래스 0)의 성능에 기인한 것이다. 즉, 모델은 클래스 0에 대해서는 잘 작동하지만, 온라인 쇼핑몰에서는 이탈율이 주요 관심사이므로 클래스 1 성능에 대해서는 심각한 성능 문제가 있다.

3.3 데이터 불균형

이에 대한 원인은 비이탈 고객이 2,894명으로 훨씬 더 많고, 반면 이탈 고객은 1,006명이어서 데이터 불균형이 존재하기 때문이다. 이럴 경우 모델 편향 현상이 발생하여 비이탈 고객에 편향되어 학습될 가능성이 있고, 이로 인해 이탈 고객을 제대로 예측하지 못하게 될 수 있다.

<표 10> 고객 분포 비중

이탈율	개수	비중
Yes	1,006	25.79%
No	2,894	74.21%

더불어 단순한 정확도 지표는 높은 비이탈율 클래스 비율 때문에 왜곡될 수 있어 평가 지표 자체에 왜곡이 발생할 수 있다. 즉, 전체 데이터를 비이탈율로 예측해도 높은 정확도를 얻을 수 있지만 실제로 마케팅에서 중요한 것은 이탈율 예측이므로 충분한 마케팅 효과를 보지 못할 것이다.

데이터 불균형 문제를 해결하기 위해 가장 일반적으로 많이 쓰이는 아래의 2가지 방법을 적용하여 모델 개선 여부를 파악하였다.

1) 언더샘플링(Under-sampling): 다수 클래스의 데이터를 줄이는 방법이나 중요한 정보를 잃을 수 있는 단점이 있다.
2) 오버샘플링(Over-sampling): 소수 클래스의 데이터를 증가시키는 방법으로 대표적인 기법으로 샘플들 사이에 새로운 샘플을 인위적으로 생성하는 SMOTE(Synthetic Minority Over-sampling Technique)가 있다.

먼저 언더샘플링(Under-sampling)을 적용하여 다수 클래스인 비이탈 고객 수를 줄여 샘플 비율을 50:50으로 맞춘 후 다시 분석하였다.

<표 11> 언더샘플링 적용 후 고객 분포

이탈율	개수	비중
Yes	1,006	50%
No	1,006	50%

<표 12> 언더샘플링 분류 모델 성능 평가 지표

Classification Report

	Precision	Recall	F1-score	Support
0	0.86	0.79	0.82	196
1	0.81	**0.87**	0.84	207
accuracy			0.83	403
macro avg	0.75	0.76	0.83	403
weighted avg	0.82	0.82	0.83	403

Confusion Matrix
[[518 78]
[64 120]]

언더 샘플링을 통해 accuracy(83%) 및 recall(87%) 성능이 개선되었음을 확인하였으며, 추가로 오버샘플링(SMOTE)도 진행하였다.

<표 13> 오버샘플링(SMOTE) 적용 후 고객 분포

이탈율	개수	비중
Yes	2,894	50%
No	2,894	50%

<표 14> 오버샘플링(SMOTE) 분류 모델 성능 평가 지표

Classification Report

	Precision	Recall	F1-score	Support
0	0.97	0.80	0.88	578
1	0.83	**0.97**	0.90	580
accuracy			0.89	1158
macro avg	0.90	0.89	0.89	1158
weighted avg	0.90	0.89	0.89	1158

Confusion Matrix
[[465 113
[16 564]]

오버샘플링은 원본 모델과 비교하여 클래스 0에 대한 recall 성능은 80%로 다소 낮아졌지만, 클래스 1에 대해서는 97%로 대폭 개선되었음을 확인하였다.

<표 15> 샘플링 기법별 Recall 값 비교

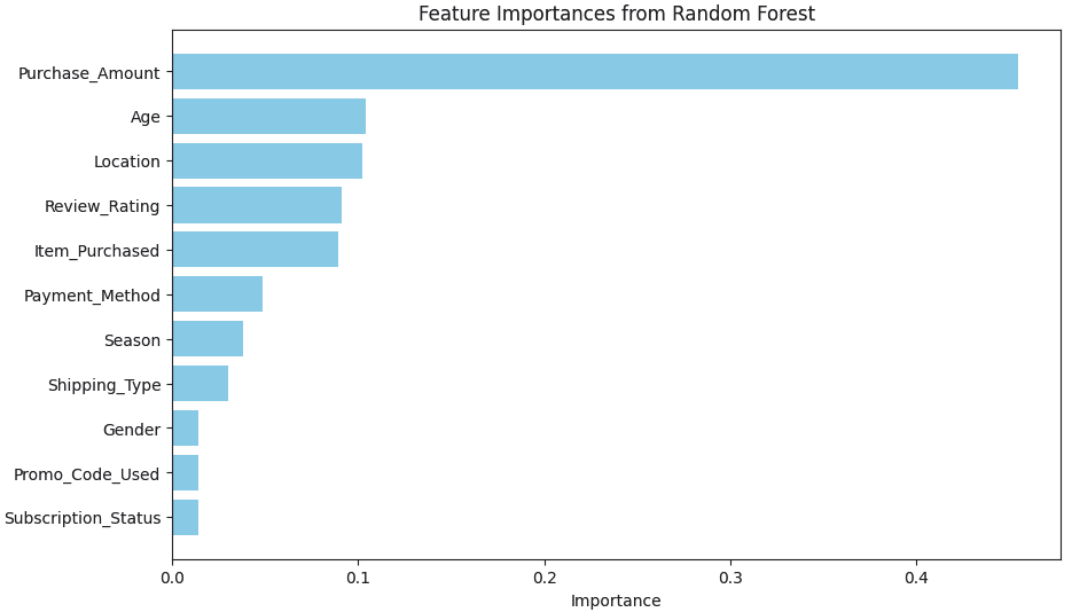
Recall	Original	Under Sampling	**Over Sampling**
Class 0 (비이탈)	0.87	0.79	0.80
Class 1 (이탈)	**0.65**	**0.87**	**0.97**

1) 데이터 균형 측면
   : Original - (클래스 0, 비이탈 고객) 596개, (클래스 1, 이탈 고객) : 184개
   : Under Sampling - (클래스 0, 비이탈 고객) : 196개, (클래스 1, 이탈 고객) 207개
   : Over Sampling(SMOTE) - (클래스 0, 비이탈 고객) : 578개, (클래스 1, 이탈 고객) 580개
   더 많은 데이터를 사용하면서도 균형을 유지하고 있기 때문에 더 좋은 성능이라 볼 수 있다.
2) 전체 정확도 측면
   : Original - 0.82(82%)
   : Under Sampling - 0.83(83%)
   : Over Sampling(SMOTE) - 0.89(89%) 훨씬 더 높은 정확도를 보인다.

결론적으로 보면 Accuracy 및 Recall 점수 모두 Over Sampling(SMOTE)이 언더 샘플링 대비해서 더 나은 성능을 보였으며 훨씬 많은 데이터(1158개 vs 403개)를 사용하면서도 더 높은 수준의 성능을 유지하고 있어 일반화 능력도 더 뛰어날 가능성이 높다. 특히, 큰 데이터셋에서 안정적인 성능을 보이고 있으므로 실제 적용 시 더욱 신뢰할 만하다고 볼 수 있다.

추가로 랜덤 포레스트(Random Forest) 모델을 사용하여 각 피처의 중요도를 파악하였다. 피처 중요도는 모델이 예측을 수행할 때 해당 피처가 얼마나 중요한 역할을 하는지를 나타낸다. 이를 통해 어떤 피처가 고객 이탈 예측에 가장 큰 영향을 미치는지 알 수 있다.

[그림 9] Random Forest 모델의 특성 중요도



Purchase_Amount(구매 금액)이 이탈 예측에서 가장 중요한 변수로 볼 수 있다. 이는 고객의 구매 금액이 높을수록 고객의 충성도가 높고 이탈 가능성이 낮음을 의미한다.

Age(나이)는 두 번째로 중요한 변수로 고객의 나이가 이탈 예측에 중요한 역할을 한다는 것을 의미한다. 특정 연령대의 고객이 다른 연령대에 비해 이탈할 가능성이 더 높거나 낮을 수 있다.

Location(위치)는 세 번째로 중요한 변수로 특정 지역에 거주하는 고객들이 다른 지역의 고객들보다 이탈 가능성이 높거나 낮을 수 있음을 의미한다.

Review_Rating(리뷰 평점)은 고객이 남긴 리뷰 평점도 중요한 변수로 나타난다. 높은 리뷰 평점을 남긴 고객들은 일반적으로 서비스에 만족하여 이탈 가능성이 낮을 수 있다.

Item_Purchased(구매한 아이템)도 어떤 아이템을 구매했는지 여부에 따라 이탈 예측에 중요한 요소가 될 수 있다.

Payment_Method(결제 방법)도 중요한 변수이며 결제 옵션의 다양성에 따라 이탈 가능성이 다르게 나타날수 있음을 의미한다.

Season(계절)도 이탈 예측에 영향을 미치며 특정 계절에 이탈율이 높거나 낮을 수 있음을 나타낸다.

Shipping_Type(배송 타입)도 중요한 변수인데 배송 옵션에 따른 배송 기간 차이로 고객들의 이탈 가능성이 있음을 나타낸다.

그 외 Gender(성별), Promo_Code_Used(프로모션 코드 사용 여부), Subscription_Status(구독 상태)는 가장 낮은 영향도를 보인다.

이 정보를 바탕으로, 기업은 중요한 변수에 집중하여 고객 이탈을 줄이기 위한 전략을 세울 수 있다. 예를 들어, 높은 구매 금액을 유지하도록 유도하거나 특정 지역 및 구매 아이템별 맞춤형 마케팅을 진행하는 것이 효과적일 수 있다.

3.4 예측 모델 성능 비교

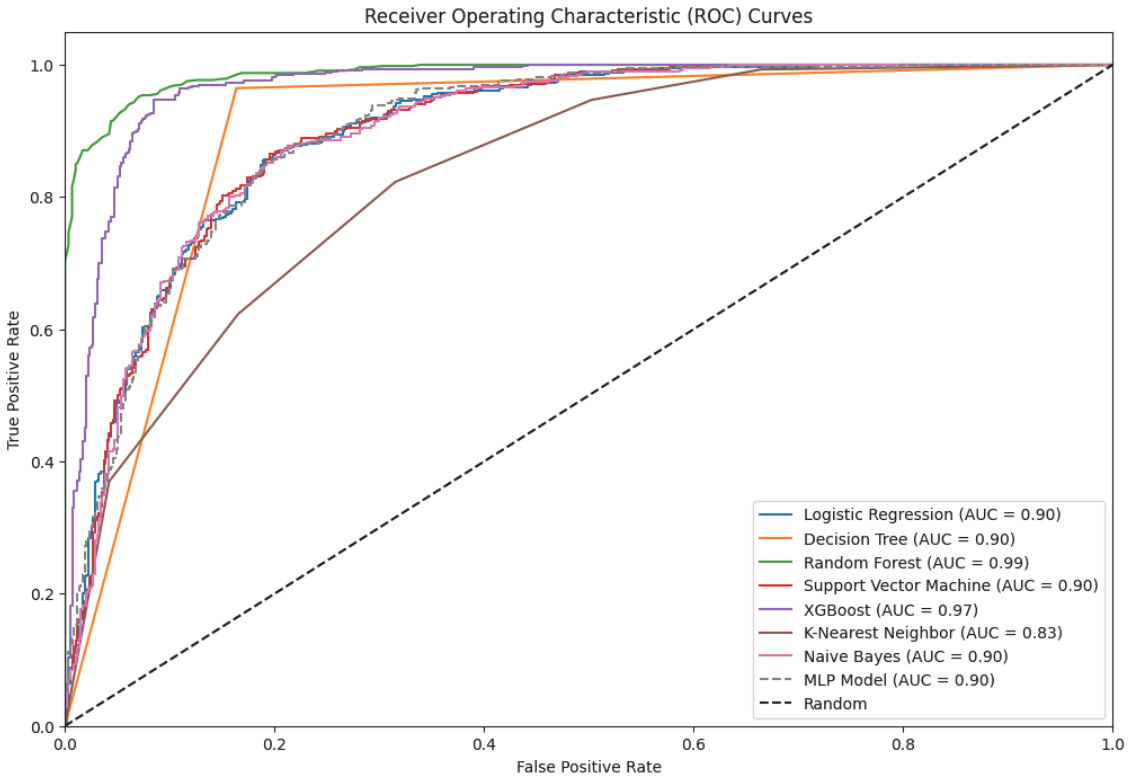
모델의 안정성을 확인하기 위해 여러 모델들의 성능 지표를 weighted 평균 기준으로 교차 검증을 수행하였다. 머신러닝 대표적인 모델 8가지인 로지스틱 회귀(Logistic Regression), 결정 트리(Decision Tree), 랜덤 포레스트(Random Forest), 서포트 벡터 머신(SVM), XGBoost, K-최근접 이웃(KNN), 나이브 베이즈(Naive Bayes), 다층 퍼셉트론(MLP)을 각 모델별 성능 지표(정확도, 정밀도, 재현율, F1점수)를 중심으로 비교하였다.

<표 16> 분류 모델별 성능 비교

	Accuracy	Precision	Recall	F1 Score
Logistic Regression	83.07%	83.40%	83.07%	83.06%
Decision Tree	89.90%	90.64%	89.90%	89.87%
Random Forest	**91.11%**	**91.91%**	**91.11%**	**91.08%**
Support Vector Machine	81.52%	82.69%	81.52%	81.41%
XGBoost	90.76%	91.48%	90.76%	90.74%
K-Nearest Neighbor	75.22%	75.89%	75.22%	75.12%
Naive Bayes	82.21%	82.86%	82.21%	82.16%
MLP	81.61%	82.40%	81.61%	81.54%

Random Forest가 모든 지표에서 균형 잡힌 성능을 보여 이탈 예측에 가장 적합한 모델로 판단된다. 반면, K-Nearest Neighbor는 가장 낮은 성능을 보인다.

[그림 10] 분류 모델별 ROC 곡선



추가로 머신 러닝 모델의 분류 성능을 시각적으로 나타내는 ROC (Receiver Operating Characteristic)를 비교하였다. 그래프의 각 곡선은 특정 모델의 분류 성능을 나타내며, 곡선 아래의 면적(AUC, Area Under the Curve) 값은 모델의 성능을 수치적으로 나타낸다.

1. X축(False Positive Rate, FPR)
   - 실제로 이탈하지 않는 고객을 이탈할 것이라고 잘못 예측하는 비율이다. 예를 들어 FPR 값이 0.1이라면 이탈하지 않는 고객의 10%를 이탈한다고 잘못 예측했음을 의미한다.
2. Y축(True Positive Rate, TPR)
   - 실제로 이탈하는 고객을 정확히 예측하는 비율이다. 예를 들어, TPR값이 0.9라면, 실제로 이탈하는 고객의 90%를 올바르게 예측했음을 의미한다.
3. 대각선(Random Classifier):
   - 대각선은 무작위 분류기의 성능을 나타내며, 무작위 추측과 동일한 성능이다.

AUC 값이 1에 가까울수록 모델의 성능이 뛰어나며 양성과 음성을 잘 구분할 수 있음을 의미한다. AUC 값이 0.5이면, 모델의 성능이 무작위 추측과 같음을 의미한다.

전체적으로 그래프를 보면 Random Forest 모델이 AUC 값 0.99로 가장 높은 성능을 보이며, XGBoost 모델도 높은 AUC 값(0.97)을 나타낸다. K-Nearest Neighbor 모델은 상대적으로 낮은 AUC 값(0.83)으로 다른 모델들에 비해 성능이 낮다. Logistic Regression, Decision Tree, SVM, Naive Bayes, MLP 모델은 모두 AUC 값이 0.90으로 유사한 성능을 보인다.

온라인 쇼핑몰에서는 일반적으로 이탈율 예측의 목표는 고객 이탈을 최소화하고, 고객 유지 전략을 개선하는데 있으므로, 재현율(Recall)이 높은 모델을 선택하는 것이 중요하다.

이탈할 가능성이 높은 고객을 최대한 많이 식별하여, 이들에게 특별한 프로모션이나 혜택을 제공함으로써 이탈을 방지하는 것이 중요하다. 비용 절감 측면에서 보면 잘못된 이탈 예측(즉, 이탈하지 않을 고객을 이탈할 것이라고 예측)으로 인한 마케팅 비용보다, 이탈할 가능성이 높은 고객을 놓치는 것(즉, 실제 이탈할 고객을 예측하지 못하는 것)으로 인한 매출 손실이 더 크므로 재현율이 가장 중요하다.

이러한 관점에서 보면, Random Forest가 가장 높은 정확도와 재현율을 보여 최종적으로 최고의 모델이라고 할 수 있다.

Ⅴ. 결론 및 시사점

1. 결과 분석 및 요약

본 연구는 미국 온라인 소매몰의 3,900명 고객 구매 데이터를 활용하여 RFM(Recency Frequency Monetary) 모델과 머신러닝 기법을 결합하여 최적의 고객 이탈 예측 모델을 개발하는 데 중점을 두었다.

1. 최적 예측 모델 개발:

- RFM(Recency, Frequency, Monetary) 모델을 기반으로 이탈 변수를 새로 정의하고 다양한 머신러닝 기법을 결합하여 분석하였다.
- 8가지 머신러닝 모델(로지스틱 회귀, 결정 트리, 랜덤 포레스트, SVM, XGBoost, KNN, 나이브 베이즈, MLP)을 비교 분석한 결과, Random Forest 모델이 91.11%의 재현율(Recall)로 가장 우수한 성능을 보였다.
- ROC 곡선 분석 결과, Random Forest 모델의 AUC 값이 0.99로 가장 높게 나타나 이탈 고객과 비이탈 고객을 매우 정확하게 구분할 수 있음을 확인하였다.

2. 주요 예측 변수 식별:
   - 구매 금액, 나이, 고객 위치, 리뷰 평점, 구매 아이템 등이 고객 이탈 예측의 주요 변수로 확인되었다.
   - 특히, 구매 금액이 가장 중요한 예측 변수로 나타났으며, 이는 고객의 지출 수준이 이탈 가능성과 밀접한 관련이 있음을 시사한다.
3. 데이터 불균형 문제 해결:
   - 초기 모델에서 나타난 데이터 불균형 문제를 해결하기 위해 언더샘플링과 오버샘플링(SMOTE) 기법을 적용하였다.
   - 이를 통해 모델 성능을 개선하였고, 특히 소수 클래스인 이탈 고객에 대한 예측 정확도를 높일 수 있었다.
4. 고객 세분화:
   - RFM 점수를 기반으로 고객을 충성 고객(38.5%), 중간 가치 고객(35.7%), 이탈 위험 고객(25.8%)으로 세분화하였다. 이는 고객 그룹별 맞춤형 전략 수립에 활용될 수 있다.
5. 실무적 시사점:
   - 높은 재현율을 보이는 모델을 통해 이탈 가능성이 높은 고객을 사전에 식별하고, 맞춤형 마케팅 전략을 수립할 수 있다.
   - 구매 금액이 주요 예측 변수임을 고려하여, 고객의 구매 금액을 증가시키는 전략(예: 교차판매, 상향판매)이 이탈 방지에 효과적일 수 있다.
   - 나이별, 지역별, 제품별로 차별화된 고객 유지 전략을 수립할 필요가 있다.
   - 리뷰 평점의 중요성을 감안할 때, 고객 만족도 향상을 위한 지속적인 노력이 필요하다.

2. 연구의 한계와 향 후 과제

사용된 데이터셋이 미국의 특정 온라인 소매몰에 한정되어 있어, 연구 결과의 일반화에 제한이 있을 수 있다. 또한, 데이터가 특정 기간에 한정되어 있어 장기적인 트렌드나 계절성을 충분히 반영하지 못했을 가능성이 있다.

분석에 사용된 변수들이 제한적이어서, 고객 행동에 영향을 미칠 수 있는 다른 중요한 요인들(예: 웹사이트 사용성, 고객 서비스 경험 등)을 고려하지 못했을 수 있다. 특히 Random Forest 모델의 경우 높은 예측 성능을 보였지만 모델의 결정 과정을 완전히 해석하기 어려운 "블랙박스" 특성이 있다.

특히, 도출된 인사이트를 바탕으로 주요 변수들의 실제 효과를 검증하지 못한 점이 주요한 한계로 지적될 수 있다. 이러한 한계점들은 향후 연구에서 더 다양한 데이터 소스의 활용, 장기적인 분석, 추가적인 변수 고려, 그리고 제안된 전략의 실제 적용 및 효과 검증 등을 통해 보완될 수 있을 것이다.

참고문헌

길기영, 서영욱 (2023). “온라인 쇼핑몰의 학습, 온라인 리뷰, 소비자 커뮤니케이션이 지각된 혜택을 통해 고객 e-충성도에 미치는 영향,” 한국산학협력학회지, 24(9), 547-561.

김관희 (2023). 머신러닝을 활용한 이커머스 환경의 고객 행동 기반 분류 연구. 석사학위논문, 중앙대학교.

김민경, 허재석, 사애진, 전아름, 이한별 (2022). “클러스터링 기법을 활용한 이커머스 사용자 리뷰에 따른 시장세분화 연구,” 전자상거래학회지, 27(2), 21-36.

박미령 (2023). “온라인 쇼핑몰에서 패션제품의 반품 요인이 성별에 따라 반품 행동과 재구매 의도에 미치는 영향“, 한국산학기술학회논문지, 24(11), 617-625.

양치복, Yangsok Kim, 이충권 (2017). “온라인쇼핑몰 구매데이터를 활용한 시계열 분석에 관한 연구,” 한국정보기술학회 추계 종합학술대회 논문집, 81-90

유희정, 도지현, 김능회 (2023). “평점에 따른 온라인 리뷰의 유용성 연구,” 한국정보기술학회 추계 종합학술대회 논문집, 978-980

최경빈, 남기환 (2019). “쇼핑 웹사이트 탐색 유형과 방문 패턴 분석,” J Intell Inform Syst, 25(1), 85-107.

Alshamsi, A. (2022). Customer churn prediction in ECommerce sector. Master’s thesis, Rochester Institute of Technology.

Bagul, N., Berad, P., Surana, P., & Khachane, C. (2021). Retail customer churn analysis using RFM model and k-means clustering. *International Journal of Engineering Research & Technology (IJERT),* 10(03), 349-354.

Bergström, S. (2019). Customer segmentation of retail chain customers using cluster analysis. Master’s thesis, Stockholm University.

Ho, T., Nguyen, S., Nguyen, H., Nguyen, N., Man, D.-S., & Le, T.-G. (2023). An extended RFM model for customer behaviour and demographic analysis in retail industry. *Business Systems Research,* 14(1), 26-53.

Mitrovic, S., Singh, G., Baesens, B., Lemahieu, W., & De Weerdt, J. (2017). Scalable RFM-enriched representation learning for churn prediction. 2017 *International Conference on Data Science and Advanced Analytics,* 79-88.

Roy, K. K., Jaiswal, P. K., & Sinha, A. (2018). Data analytics for consumer purchasing pattern with behavioral segmentation. 3*rd International Conference on Internet of Things and Connected Technologies* (ICIoTCT).

Xu, T., Ma, Y., Ao, C., Qu, M., & Meng, X. (2023). A novel telecom customer churn analysis system based on RFM model and feature importance ranking. *Interdisciplinary Journal of Information, Knowledge, and Management,* 18, 719-737.

Zuo, Y., Ali, A. B. M. S., & Yada, K. (2014). Consumer purchasing behavior extraction using statistical learning theory. *Procedia Computer Science,* 35, 1464-1473.

Abstract

Churn Prediction Using Machine Learning based on E-commerce Data

In Jik Lee

The main objective of this study is to accurately predict customer churn in order to improve customer retention strategies and enhance long-term profitability for businesses.

The analysis was conducted by combining the RFM (Recency, Frequency, Monetary) model with various machine learning techniques, based on purchase data from 3,900 customers of an online retail store. The results showed that the Random Forest model demonstrated the best predictive performance with a recall rate of 91.11%. Purchase amount, age, customer location, review ratings, and purchased items were identified as key variables for predicting customer churn. Notably, purchase amount emerged as the most important predictive variable, suggesting a close relationship between customer spending levels and the likelihood of churn. To address the issue of data imbalance, undersampling and oversampling (SMOTE) techniques were applied to improve model performance, thereby increasing the prediction accuracy for the minority class of churning customers.

Customers were segmented into loyal customers (38.5%), medium-value customers (35.7%), and high-risk churn customers (25.8%) based on their RFM scores. This segmentation can help in providing special promotions or personalized services to high-risk customer groups to reduce churn rates.

Comparing various machine learning models, Random Forest outperformed in all metrics including accuracy, precision, recall, and F1 score, indicating its ability to capture complex customer behavior patterns effectively. Moreover, ROC curve analysis showed an AUC value of 0.99, confirming the model's high accuracy in distinguishing between churning and non-churning customers.

These research findings can be valuable for online retailers in predicting customer churn and developing effective customer retention strategies. Particularly, the high-recall model allows for early identification of customers with high churn probability, enabling the development of tailored marketing strategies. This can lead to improved customer retention rates and ultimately contribute to increased business profitability.

Future research should focus on applying a wider range of variables and cutting-edge algorithms to improve model accuracy, as well as analyzing real-world business case studies to validate the model's effectiveness. Through these efforts, it is expected that the data-driven decision-making capabilities in the online retail industry can be further enhanced, contributing to the development of customer-centric business strategies.

Keywords: Customer Churn Prediction, Machine Learning, E-commerce, Online Shopping, RFM Model, Random Forest, Data Imbalance, Exploratory Data Analysis, Customer Segmentation