석사학위논문

이차원 고객충성도 세그먼트 기반의 고객이탈예측 방법론

2021년

한 성 대 학 교 대 학 원

산 업 경 영 공 학 과

산 업 경 영 공 학 전 공

홍 승 우

석사학위논문
지도교수 김형수

이차원 고객충성도 세그먼트 기반의 고객이탈예측 방법론

A Methodology of Customer Churn Prediction based on Two-Dimensional Loyalty Segmentation

2020년 12월 일

한 성 대 학 교 대 학 원

산 업 경 영 공 학 과

산 업 경 영 공 학 전 공

홍 승 우

석사학위논문
지도교수 김형수

이차원 고객충성도 세그먼트 기반의 고객이탈예측 방법론

A Methodology of Customer Churn Prediction based on Two-Dimensional Loyalty Segmentation

위 논문을 공학 석사학위 논문으로 제출함

2020년 12월 일

한 성 대 학 교 대 학 원

산 업 경 영 공 학 과

산 업 경 영 공 학 전 공

홍 승 우

홍승우의 공학 석사학위 논문을 인준함

2020년 12월 일

심사위원장 _____(인)

심 사 위 원 _____(인)

심 사 위 원 _____(인)

국 문 초 록

이차원 고객충성도 세그먼트 기반의 고객이탈예측 방법론

한 성 대 학 교 대 학 원
산 업 경 영 공 학 과
산 업 경 영 공 학 전 공
홍 승 우

CRM의 하위 연구 분야로 진행되었던 고객이탈예측은 최근 비즈니스 머신러닝 기술의 발전으로 인해 빅 데이터 기반의 퍼포먼스 마케팅 주제로 더욱 그 중요도가 높아지고 있다. 그러나, 기존의 관련 연구는 예측 모형 자체의 성능을 개선시키는 것이 주요 목적이었으며, 전체적인 고객이탈예측 프로세스를 개선하고자 하는 연구는 상대적으로 부족했다. 본 연구는 성공적인 고객이탈관리가 모형 자체의 성능보다는 전체 프로세스의 개선을 통해 더 잘 이루어질 수 있다는 가정 하에, 이차원 고객충성도 세그먼트 기반의 고객이탈예측 프로세스 (CCP/2DL: Customer Churn Prediction based on Two-Dimensional Loyalty segmentation)를 제안한다. CCP/2DL은 양방향, 즉 양적 및 질적 로열티 기반의 고객세분화를 시행하고, 고객세그먼트들을 이탈패턴에 따라 2차 그룹핑을 실시한 뒤, 이탈패턴 그룹별 이질적인 이탈예측 모형을 독립적으로 적용하는 일련의 이탈예측 프로세스이다. 제안한 이탈예측

프로세스의 상대적 우수성을 평가하기 위해 기존의 범용이탈예측 프로세스와 클러스터링 기반 이탈예측 프로세스와의 성능 비교를 수행하였다. 글로벌 NGO 단체인 W사의 협력으로 후원자 데이터를 활용한 분석과 검증을 수행했으며, 제안한 CCP/2DL의 성능이 다른 이탈예측 방법론에 비해 더 우수한 성능을 보이는 것으로 나타났다. 이러한 이탈예측 프로세스는 이탈예측에도 효과적일 뿐만 아니라, 다양한 고객통찰력을 확보하고, 관련된 다른 퍼포먼스 마케팅 활동을 수행할 수 있는 전략적 기반이 될 수 있다는 점에서 연구의 의의를 찾을 수 있다.

[주요어] 고객이탈예측, CRM, 고객관계관리, 고객로열티, 고객 빅데이터

목 차

표 목 차

그 림 목 차

제 1 장 서 론

제 1 절 연구의 배경

성숙기에 접어든 대부분의 업종들이 치열한 경쟁환경에 노출되면서 고객 생애가치의 중요성을 인식하고, 이에 따라 신규고객의 확보보다는 기존고객의 이탈방지가 더욱 중요한 비즈니스 이슈가 되고 있다(Hung & Tsai, 2008). 이는 기존 고객을 유지하는 것이 새로운 고객을 확보하는 것에 비해 경제적인 측면에서 훨씬 유리하기 때문인데(Nie, Wang, Zhang, Tian, & Shi, 2009), 실제로 신규 고객의 획득 비용은 기존 고객의 유지비용 대비 5~6배에 달하는 것으로 알려져 있다(Athanassopoulos, 2000). 또한, 이탈 고객은 기업의 안정적인 매출을 급격하게 감소시키기 때문에 기존 매출을 유지시키기 위해서는 일정 수준의 신규 고객을 지속적으로 유치해야 하는 비효율적인 경영구조에 빠지게 한다(김형수, 신동엽, 2012). 이와는 반대로 고객이탈을 효과적으로 방지하여 고객 유지율을 개선시키는 기업은 회사의 수익성 증대뿐만 아니라, 고객만족도 향상을 통해 브랜드 이미지 개선에도 긍정적인 효과를 가져오는 것으로 알려져 있다(Hung, Yen, & Wang, 2006). 즉, 현재와 같은 무한경쟁 환경 속에서는 고객의 이탈을 효과적으로 예측하고, 방지하는 것이 재무적, 비재무적 관점에서 기업에게 효과적이다.

그동안의 고객이탈예측 연구는 이동통신업, 금융업, 유통업, 그리고 게임 산업 등 경쟁이 치열하고, 이탈관리가 시급한 업종에서 활발히 진행되어 왔으며, 주로 머신러닝과 인공지능 기술을 활용한 예측모형 개발에 초점이 맞추어져 왔다. 이러한 기존의 이탈예측 연구는 단순히 다양한 모형의 성능을 비교하거나(김충영, 장남식, 김준우, 2003), 이탈예측에 효과적인 변수(Feature)를 탐색하거나(De Bock & Van den Poel, 2012), 새로운 앙상블 기법을 개발하는 것(이재식, 이진천, 2006)과 같이 이탈예측 모형 자체의 성능을 개선하는 것에 집중되었고, 대부분의 연구에서 전체 고객집단을 하나의 그룹으로 간주하고 예측모형을 개발했기 때문에 실질적인 활용도 측면에서 한계를 가진다. 실제 비즈니스 상에서의 고객들은 이질적인 거래패턴으로 인

해 고객의 행태 특성이 다르고, 이에 따른 이탈률도 다르기 때문에 고객 전체를 하나의 고객집단으로 전제하는 것은 무리가 따르기 때문이다(김형수, 신동엽, 2012). 이처럼 전체 고객을 대상으로 단일 이탈예측 모형이 적용될 경우, 특정 고객 세그먼트의 고객행태 특성이 달라질 때 변화된 고객특성이 이탈예측 모형에 반영되기 어렵기 때문에 이러한 방법론은 개인별 기여가치가 이질적인 업종에서 효과적인 고객이탈예측 성과를 기대하기 어려워진다.

따라서, 고객의 행태특성이 이질적인 업종에서 효과적인 고객이탈 예측을 수행하기 위해서는 충성도와 같은 고객 분류 기준에 따라 고객을 세분화하고, 이를 기반으로 적합한 이탈예측 모형이 개별적으로 운영되는 것이 바람직하다. 물론 일부 연구에서는 클러스터링 기법으로 고객을 세분화하여 개별 고객그룹에 대한 이탈예측 모형을 적용한 연구가 존재한다(오세준, 이은조, 우지영, 김휘강, 2018). 이러한 이탈예측 프로세스는 전체 고객집단을 대상으로 한 단일 이탈예측 모형에 비해 우수한 예측결과를 가져올 수 있지만, 클러스터링은 투입변수들을 바탕으로 거리를 계산하는 기계적이고, 탐색적인 그룹핑 기법이기 때문에 로열티와 같은 기업의 전략적 의도가 제대로 반영되지 못한다는 점에서 여전히 개선의 여지가 남아있다.

이에 본 연구에서는 비즈니스 머신러닝의 성과가 알고리즘이나 모형 자체의 우수성보다는 전체 머신러닝 프로세스의 우수성에 의해 더 큰 영향을 받는다는 믿음을 기반으로 이차원 고객충성도 세그먼트 기반의 고객이탈예측 프로세스(CCP/2DL:Customer Churn Prediction based on Two-Dimensional Loyalty segmentation)를 제안하고자 한다.

제 2 장 문헌 연구

제 1 절 머신러닝 기반 CRM 연구

1) 머신러닝

가) 머신러닝 개념

머신러닝(Machine Learning)이란 데이터로터 지식을 습득하는 과정을 컴퓨터를 통해 자동화하는 방법론을 의미한다(Langley & Simon, 1995). 머신러닝과 관련된 연구들은 기존에 기술적 한계로 인해 정체되어 왔으나 최근 등장한 빅데이터 기술 분야에서 점점 실현 가능성이 커지고 있다(최도현, 박중오, 2015). 일반적으로 머신러닝은 알고리즘 성격에 따라 크게 지도 학습(Supervised Learnig)과 비지도 학습(Unsupervised Learning)으로 분류한다.

지도학습 모형이란 명시적인 지도를 통해 모형이 학습하는 방법으로 학습 데이터에 이미 정답이 주어진 상태에서 시스템을 학습시키는 방법을 말한다. 지도학습 머신러닝 모형은 크게 대상을 목적에 따라 구분하는 분류(Classification) 모형과 특정 값을 예측하는 수치 예측(Regression, 회귀)모형으로 구분할 수 있다. 분류는 입력 데이터들, 즉 대상을 목적에 따라 특정 그룹으로 구분하는 적업을 의미한다. 반면, 수치예측 모형에 결과 값은 특정 함수식으로 계산된 임의 값으로 이를 계산하는 작업이 수치 예측에 해당한다(김형수, 2020).

비지도 학습은 명시적인 지도 없이 모형이 학습하는 방법으로 학습 데이터에는 정답 레이블에 해당하는 종속변수(결과변수)가 존재하지 않아 모형이 학습하는 과정에서 어떠한 지도 메커니즘도 발생하지 않는다. 비지도학습 모형은 주로 군집발견, 유사 패턴파악, 상품추천, 의미 추출 등 매우 다양한 분석 목적에 활용되며, 기본적으로 분류나 예측과 같이 지도학습 메커니즘이 존재하지 않는 모든 머신러닝 모형이 비지도 학습 모형이라고 볼 수 있다(김형수, 2020).

나) 머신러닝 성능 평가

머신러닝의 성능 평가 방법 또한 머신러닝 모형에 따라 달라진다. 일반적으로 지도학습 모형 중에서도 분류 모형의 경우 정확도, 정밀도, 재현율, F1-스코어 등을 활용하여 성능을 평가하며, 수치 예측 모형에 경우에는 결정 계수, RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) 등과 같이 실제값과 예측값의 차이를 기반으로 하는 지표들을 사용하여 모형의 성능을 평가한다. 반면, 비지도학습 모형의 경우는 지도학습처럼 명확한 정답이 존재하지 않고 탐색적으로 결론을 유도하는 목적을 갖고 있으므로 모형의 성능을 평가한다는 개념 자체가 성립되기 어렵다(김형수, 2020). 다만, 군집분석과 협업필터링의 경우에는 사후적으로 결과를 살펴보는 성과평가 방식으로 비지모형의 성능을 평가해볼 수 있다. 본 연구에서 사용한 지도학습 모형의 성능평가는 분류모형의 성능평가 방식인 정확도, 정밀도, 재현율, F1-스코어를 사용하였으며 위 지표들은 아래 〈표 2-1〉에 나온 것과 같이 혼동행렬을 기반으로 계산한다.

〈표 2-1〉 혼동행렬

혼동행렬		예측 결과	
		양성(Positive)	음성(Negative)
실제 결과	양성(Positive)	참 양성(TP)	거짓 음성(FN)
	음성(Negative)	거짓 양성(FP)	참 음성(TN)

가) 정확도(Accuracy)

정확도란 전체 중에서 올바르게 예측한 것이 몇 개인지를 살펴보는 지표로 계산식은 아래와 같다.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

나) 정밀도(Precision)

정밀도란 모형이 양성이라고 예측한 것 중 실제로 양성이 얼마나 되는지를 판단하는 지표이며 계산식은 아래와 같다.

$$Precision = \frac{TP}{TP + FP}$$

다) 재현율(Recall)

재현율이란 실제 양성 데이터 중에서 모형이 양성으로 분류한 비율을 나타내는 지표이며, 민감도(Sensitivity)라고 표현하기도 한다. 재현율의 계산식은 아래와 같다.

$$Recall = \frac{TP}{TP + FN}$$

라) F1-스코어(F1-Score)

F1-스코어는 정밀도와 재현율이 갖는 상충관계를 고려하여 두 가지 지표를 모두 반영한 모형성능 평가지표이며 계산식은 아래와 같다.

$$F1 - Score = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

정밀도와 재현율은 모형의 실질적인 성능을 평가하는 주요 지표이나 두

지표 사이에는 어느 정도 상충관계(Trade-Off Relationship)가 존재한다. 즉, 정밀도가 높아지면 재현율이 낮아지고 재현율이 낮아지면 정밀도가 올라가는 패턴을 보인다. 따라서 보다 정확한 모형의 성능을 평가하기 위해서는 상황에 맞게 정확도와 정밀도 평가지표 고려해야 하며 동시에 이 두 평가지표를 조화 평균하여 계산한 F1-Score를 활용할 수 있다.

2) 머신러닝을 활용한 CRM 분야

머신러닝은 비즈니스 분야 중 CRM과 같은 고객 데이터 분석 주제에 많이 활용되어 왔으며(김형수, 박예린, 이정민, 2019), 특히, CRM 분야 중에서도 아래 〈표 2-2〉에서 정리한 고객세분화, 고객이탈예측, 그리고 개인화 추천 분야에서의 머신러닝 기반 CRM 연구가 다른 분야들에 비해 활발히 이루어져 왔다.

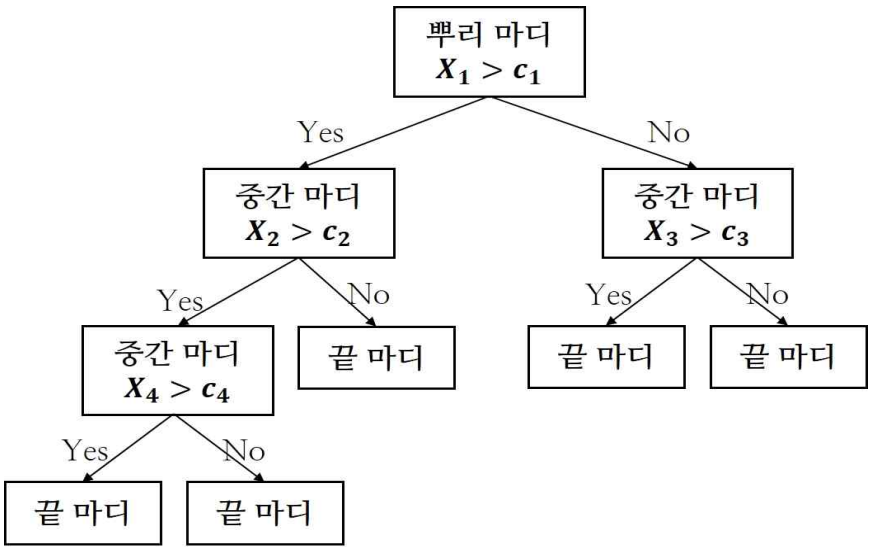
가) 고객세분화

CRM 활동의 출발점이라고 할 수 있는 고객세분화는 동질적인 고객의 특성을 중심으로 고객을 그룹핑하는 활동으로서 고객그룹별 차별화된 CRM 활동의 기반을 제공한다(장민석, 김형중, 2018). 고객세분화를 통해 기업은 다양한 판매와 홍보 활동을 저비용으로 실행할 수 있으며, 불특정 다수의 고객을 대상으로 하는 경우보다 높은 성과와 효율을 기대할 수 있게 된다(정담희, 안가경, 2018). 이와 같이 고객 세분화는 불필요한 비용을 줄이고 우수한 고객과의 지속적인 관계를 통해 수익을 창출할 수 있다(정미월, 2008). 고객세분화를 위해 사용된 머신러닝 모형은 의사결정나무와 같은 지도학습 모형이나 SOM (Self-Organizing Map) 또는 K-means 모형과 같은 비지도 학습모형이 주로 사용되었다(장남식, 2005). 최근 머신러닝 기반 고객세분화 연구의 주요 특징 중 하나는 단순히 의미 있는 고객 세그먼트를 도출하는 것에서 끝나는 것이 아니라, 고객이탈예측과 같은 연관된 다른 CRM 연구 목적을 위해 사전적인 고객세분화가 수행되고 있다는 점이다(이건창, 권순재, 신경식, 2001; Tsai & Lu, 2009; Xie, Li, Ngai, & Ying, 2009; 김담

희, 안가경, 2018; 오세준, 이은조, 우지영, 김휘강, 2018).

(1) 의사결정나무

의사결정나무(Decision Tree) 또는 나무모형은 가장 널리 알려진 머신러닝 알고리즘의 하나로 인공지능 분야 등 다양한 분야에서 분류 예측을 위해 활용되어 진다(Elouedi, Mellouli, & Smets, 2000). 의사결정나무는 의사결정 규칙을 나무구조로 표현하여 전체 자료를 몇 개의 소집단으로 분류하는 방법이며 그 구조는 아래 〈그림 2-1〉과 같다. 기본적으로 뿌리 마디(Root Node), 중간 마디(Internal Node), 끝 마디(Terminal Node), 가지(Branch)로 구성되며 상위마디가 하위마디로 분기될 때, 상위마디는 부모 마디(Parent node), 하위마디는 자식마디(Child node)라 한다. 의사결정나무는 뿌리나무에서 시작해 자식마디로 분기되고, 분기된 마디는 다시 부모마디가 되어 자식마디로 분기되는 과정이 반복된다. 이 과정에서 의사결정나무는 자식마디의 불순도가 부모마디의 불순도보다 낮게 나오는 방향으로 분기되며 불순도 계산은 카이제곱 통계량의 p-값, 지니 지수(Gini Index), 엔트로피 지수(Entropy Index)가 사용된다. 의사결정나무는 목표변수의 데이터 타입에 따라 범주형인 경우 분류나무(Classification Tree), 수치형 타입에 경우 회귀나무(Regression Tree)로 분류한다.

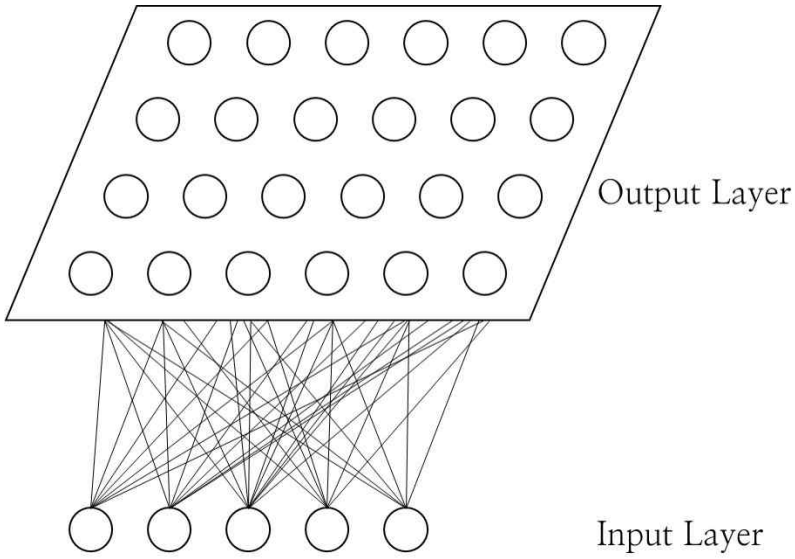


〈그림 2-1〉 의사결정나무 기본 구조

(2) SOM (Self-Organizing Map)

SOM (Self-Organizing Map)은 인공신경망(ANN)의 한 유형으로 클러스터링 기법 중 K-means와 더불어 가장 많이 사용되고 있는 방법이다 (Duda, Hart, & Stork, 2012; Tsai & Hung, 2014). 차원축소(Dimensionality Reduction)와 군집화(Clustering)를 수행할 수 있으며 외부의 피드백이나 지도 없이 스스로 학습하여 입력자료에서 의미있는 패턴이나 특징을 발견한다(Verdu. S. V., 2006). SOM은 고차원의 데이터를 사람의 눈으로 확인할 수 있는 저차원(2차원 또는 3차원)의 뉴런으로 정렬하여 지도의 형태로 형상화하며, 그 구조는 크게 입력벡터를 받는 입력층(Input Layer)과 출력층(Output Layer) 두 개의 층으로 구분할 수 있다. 입력층에 입력된 자료는 반복계산을 거쳐 입력자료를 최적화하고 자료의 차이성에 기인하여 분류하는 훈련의 과정을 거치게 된다. 학습의 경우 사전에 조사된 자료를 인공신경망에 반복 입력하여 입력층과 출력층 사이의 연결계수(Connectivity Weight)가 입력 자료의 정보 특성을 반영하도록 하는 과정이며 최종 결과물은 분석결과의 효율적인 시각화를 위해 이차원격자에 N개의 출력 뉴런으로 나타나며, 가로 및 세로 방향에 대한 최적의 배열을 위해 육

각형의 격자를 사용한다. SOM은 고차원의 데이터를 저차원의 지도 형태로 형상화하기 때문에 시각적으로 이해하기 쉽고, 입력 변수의 위치 관계를 그대로 보존하고 있어 실제 데이터가 유사하면 지도상에도 유사하게 표현할 수 있다. 또한, 역전파(Back Propagation) 알고리즘 등을 이용하는 인공신경망과 달리 하나의 전방 패스(Feed-Forward Flow)를 사용함으로써 분석 속도가 빠르다. 하지만 수학 연산상의 복잡성으로 인해 수천 개 이상의 데이터 세트는 분석하는 것이 어렵다는 단점도 존재한다.

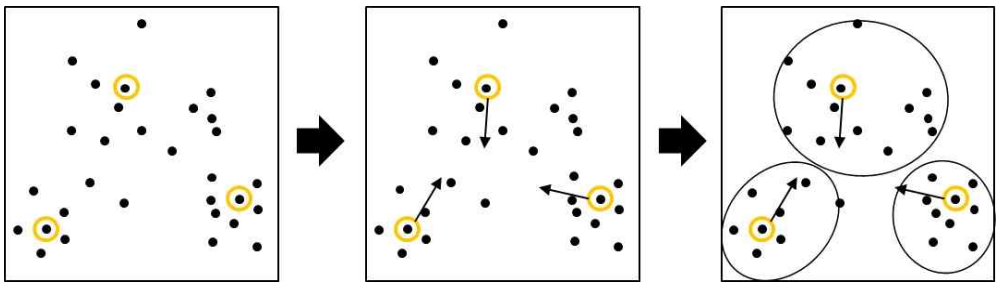


〈그림 2-2〉 SOM(Self-Organizing Map) 기본 구조

(3) K-means 클러스터링

클러스터링 기법 중 SOM과 더불어 가장 많이 사용되고 있는 방법 중 하나인 K-means 클러스터링은 비계층적 클러스터링 기법에 해당한다. K-means 클러스터링은 군집의 수를 분석가가 미리 설정한 상태에서 시행하는데 군집의 수가 K로 지정되면 군집별로 초기 중심(Initial Centroid)이 할당되며 각 관측치는 자신과 가장 가까운 중심을 기준으로 군집화 한다.

K-means 군집분석의 자세한 프로세스는 〈그림 2-3〉과 같다.



〈그림 2-3〉 K-means 클러스터링 프로세스

K-means 클러스터링 프로세스는 다음과 같다.

① 지정된 K 수 만큼 각 군집의 초기 중심을 무작위로 지정

② 각 데이터들과 초기 중심 사이의 거리를 측정하여 군집을 할당

③ 각 군집의 중심을 재설정하고, 이를 중심으로 거리 측정과 군집 할당을 다시 한다.

④ 중심이 변하지 않을 때까지 중심 재설정, 거리 측정, 군집 할당을 반복한다.

K-means 클러스터링은 탐색적인 기법으로 데이터 내부구조에 대한 사전정보 없이도 의미 있는 자료구조를 찾아낼 수 있으며 비교적 간단한 알고리즘으로 대규모의 데이터에서도 사용이 가능하다는 장점을 가지고 있다.

나) 고객이탈예측

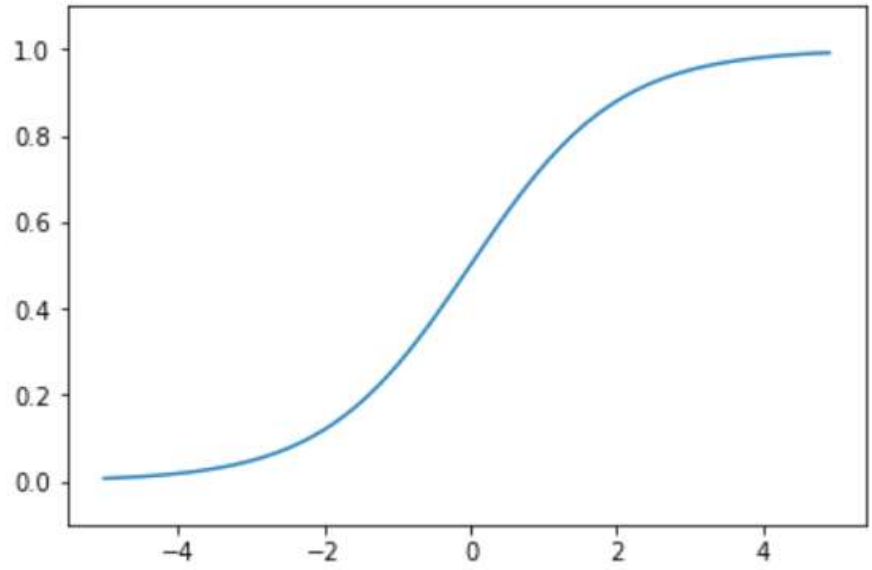
고객이탈예측 역시 머신러닝 기반의 주요 CRM 연구 주제 중 하나이다(김형수, 신동엽, 2012). 경쟁이 치열한 현대 경영 환경 하에서 고객의 이탈이 증가함에 따라 효과적인 이탈 예측은 CRM뿐만 아니라 전사적인 경영전략 차원에서도 매우 중요한 연구 주제로 인식됨에 따라(Athanassopoulos, 2000), 고객이탈을 성공적으로 예측하기 위한 새로운 모형개발 연구가 많이 이루어져 왔다(김형수, 신동엽, 2012; 오세준, 이은조, 우지영, 김휘강, 2018). 과거에는 고객의 이탈을 예측하기 위해 의사결정나무, 로지스틱 회

귀, 인공신경망과 같은 단일 알고리즘을 사용해 모형을 학습하고 성능을 비교 평가하려는 연구가 주를 이루었다면(Hung, Yen, & Wang, 2006), 최근에는 다양한 알고리즘을 결합한 앙상블 모형이나 단계적으로 이종 모형을 연결한 하이브리드 모형을 개발하려는 시도가 많아졌다(오세준, 이은조, 우지영, 김휘강, 2018; Tsai & Lu 2009).

(1) 로지스틱 회귀

로지스틱 회귀모형은 이분형 분류무형 중 가장 널리 사용되고 있는 모형으로, 통계학 및 데이터마이닝을 위주로 비즈니스, 금융, 범죄학, 공학, 의료 등 많은 분야에서 활발히 연구되어 왔다(Menard, 2002). 로지스틱 회귀분석은 아래 〈그림 2-4〉와 같은 특징을 가지는 로지스틱 함수(Logistic Function)를 사용해 특정 집단 또는 범주에 속할 확률값을 추정하여 분류 예측하는 모형으로써 범주형 혹은 명목형 지표로 구성된 종속변수에 대해서만 사용이 가능한 알고리즘이다. 로지스틱 회귀분석은 두 개의 집단을 분류하는 이항 로지스틱 회귀분석과 세 개 이상의 집단을 분류하는 다항 로지스틱 회귀 분석으로 구분한다.

$$Logistic\,function = \frac{1}{1+e^{-(\beta_0+\beta_1\chi_1+\beta_2\chi_2+\cdots\beta_p\chi_p}}$$

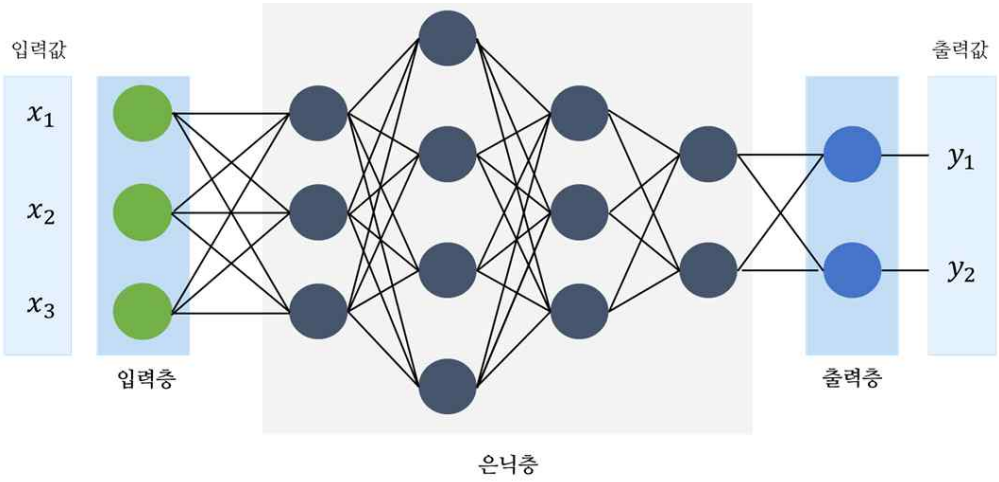


〈그림 2-4〉 로지스틱 함수

(2) 인공신경망

인공신경망은 인간의 신경세포 뉴런(Neuron)과 각 뉴런을 연결하는 시냅스(Synapse)로 구성된 인간의 뇌 신경망을 노드(Node)와 링크(Link)로 재구현한 학습 알고리즘으로, 1980년대에 컴퓨터의 처리 속도가 인공신경망의 요구를 충족시키기 시작하면서(Cochocki & Unbehauen, 1993) 활용되기 시작하였다. 인공신경망은 〈그림 2-5〉와 같이 입력 데이터를 받는 입력층(Input Layer), 입력층으로부터 받은 입력값을 계산하는 은닉층(Hidden Layer), 처리된 결과를 계산하고 출력하는 출력층(Output Layer)으로 구성된 다층 퍼셉트론(Multi-Layer Perceptron)의 구조를 가진다. 입력층의 경우 입력변수와 1:1로 대응되므로 입력층의 개수는 입력변수의 개수와 동일하다. 반면 출력층은 종속변수의 데이터 타입에 의해 노드의 개수가 결정된다. 마지막 은닉층의 경우는 입력층과 출력층과는 달리 하나 이상의 층을 가질 수 있으며 은닉층의 개수가 많아져 깊어진 인공신경망을 딥러닝(Deep learning)이라한다. 인공신경망은 수치예측, 분류예측 모두 가능하다는 점에서 그 활용도가 높으며 복잡한 데이터 집합에 패턴을 찾아내는 능력도 뛰어나다. 또

한 정보가 많은 노드에 분산되어 있기 때문에 일부 노드에 문제가 발생하여도 전체 시스템에 큰 문제를 발생시키지 않는다는 장점이 있다. 다만, 블랙박스 모형으로 결과가 나오기까지 중간 과정을 파악하기 어려우며 복잡한 학습 과정으로 인해 최적의 모형을 구축하기까지 많은 시간이 소요된다는 단점이 있다.

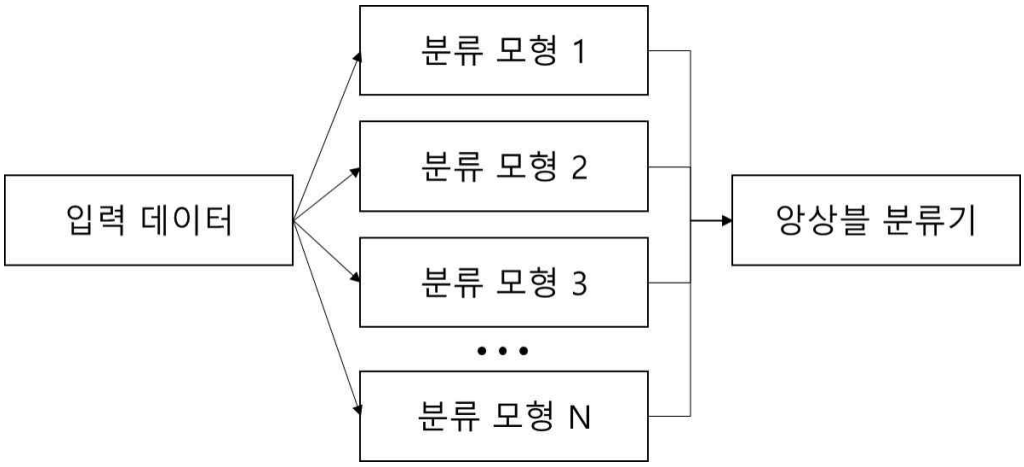


〈그림 2-5〉 인공신경망 기본 구조

(3) 앙상블 모형

앙상블 모형은 비교적 성능이 낮은 예측모형 여러 개를 결합하여 보다 좋은 성능을 보이는 예측모형을 생성하는 머신러닝 기법을 말한다(Zaïane, 1999). 그 기본 구조는 아래 〈그림 2-6〉과 같으며 앙상블의 성능은 다양한 연구를 통해서 기존의 단일 분류 모형에 비해 우수하다는 사실이 여러 연구들을 통해서 증명되었다. 앙상블은 학습 유형에 따라 분류되는데 가장 많이 사용되는 학습 유형은 배깅(Bagging), 보팅(Voting), 부스팅(Boosting) 방식이다. 배깅은 Bootstrap Aggregation의 약자로 데이터를 부트스트랩(Bootstrap) 기법을 활용해 여러 번 추출하여 각 모형에 학습시키고 학습의 결과를 집계(Aggregation)하는 방법이다. 가장 대표적인 모형으로는 랜덤 포

레스트(Random Forest)가 있다. 보팅은 배깅과 다르게 하나의 데이터 세트로 여러 모형을 학습시킨 후 각 모형의 결과를 투표 방식으로 산출하여 최종 예측하게 하는 방법이다. 보팅은 투표 방식에 따라 하드 보팅과 소프트 보팅 방식으로 구분된다. 마지막 부스팅 방식은 단일 모형을 이용하여 여러 번의 학습을 진행하여 오류를 점차적으로 줄여가는 방식으로 진행한다. 부스팅의 가장 대표적인 모형으로는 그래디언트 부스팅(Gradient Boosting), 에다부스트(AdaBoost), XGBoost 등이 있다.



〈그림 2-6〉 앙상블 모델 기반의 분류모형 구조

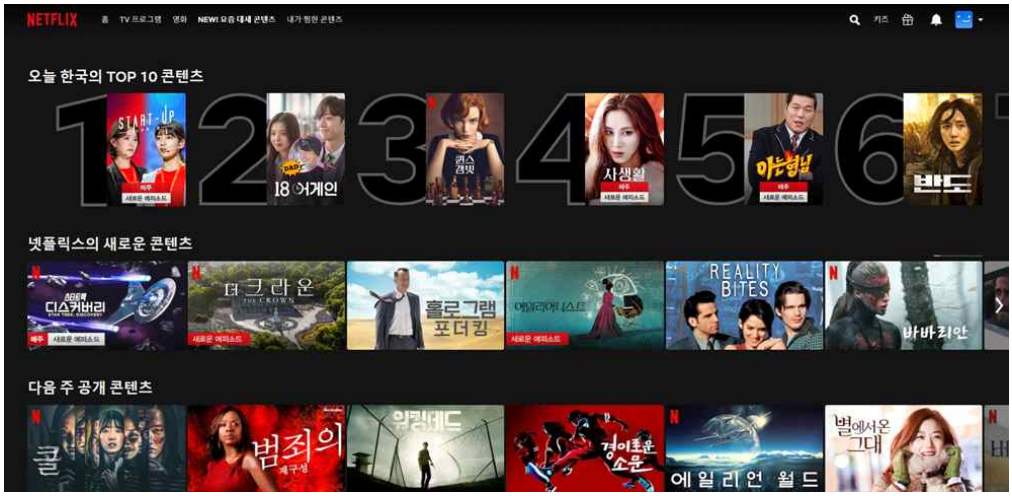
다) 개인화 추천

한편, 개인화 추천 역시 이탈예측과 더불어 가장 활발하게 진행되어 온 머신러닝 기반 CRM 연구 주제 중 하나이다(박성혁, 김형수, 2012; Wen & Zhou, 2012). 과거에는 전형적인 비지도 학습모형인 연관성 분석이나 개별 상품에 대한 구매확률 추정을 통한 개인화 추천 연구가 주를 이루었다면(조영성, 허문행, 류근호, 2008; 박성혁, 김형수, 2012), 최근에는 아마존이나 넷플릭스 등의 추천서비스에 적용된 협업필터링 기법을 활용한 개인화 추천

연구가 늘어나고 있는 추세이다. 협업필터링에 의한 개인화 추천 연구도 이탈예측 연구와 마찬가지로 예측 성능 자체를 높이기 위한 모형개발 연구(Pham, Cao, Klamma, & Jarke, 2011; Wen & Zhou, 2012)가 주를 이루었으나, 최근에는 협업 필터링에 내재된 한계점을 극복하기 위해 다양한 보조적인 처리기술을 결합한 하이브리드 협업 필터링 방법론에 대한 연구가 활발해지고 있다(조윤호, 방정혜, 2011; 조용민, 남기환, 2017).

(1) 협업 필터링(Collaborative Filtering)

추천 기술의 종류는 협업(Collaborative), 내용 기반(Content-based), 인구 통계 기반(Demographic), 효용 기반(Utility-based)과 같이 다양하다. 이 중 협업 필터링은 최근 개인화 추천에 연구와 실무에서 가장 많이 사용되고 있는 기법이다. 협업 필터링은 기존 고객들이 상품에 평가한 점수를 기반으로 하는 추천 기법으로, 해당 사용자와 비슷한 성향의 사용자들이 선호하는 아이템을 추천하는 일련의 기법을 의미한다. 협업 필터링은 크게 사용자나 상품을 기반으로 평점의 유사성을 이용하는 메모리 기반 협업 필터링(Memory based Collaborative Filtering)과 머신러닝 모형을 활용하는 모델 기반 협업필터링(Model-based Collaborative Filtering)으로 구분할 수 있으며 메모리 기반 협업 필터링은 다시 사용자 기반 협업 필터링(User-Based Collaborative Filtering)과 아이템 기반 협업 필터링(Item-Based Collaborative Filtering)으로 구분할 수 있다. 협업 필터링은 결과가 직접적이며 상품 자체에 대한 정보 없이도 추천이 가능하다는 장점이 있지만 새로운 항목 추천에는 콜드 스타트 문제가 있어 추천하기 어렵다.



〈그림 2-7〉 넷플릭스 추천 서비스 화면

〈표 2-2〉 머신러닝 기반 CRM 연구

분류		연구	주요 내용
고객 세분화	단일 모형	홍태호, 서보밀, 2004	고객의 심리통계학적 데이터에 k-means를 적용하고 판별분석과 인공신경망을 이용해 고객세분화 집단 예측
		장남식, 2005	사전 세분화를 통해 선정된 고객집단을 중심으로 관리하는 것이 중요하다는 사실을 증명
	하이브리드 모형	Seret, Verbraken, Versailles, Baesens , 2012	프로필 생성을 위한 새로운 SOM 기반 방법: 직접 마케팅의 이론과 적용
		Liao, Chen, Liu, & Chiu, 2015	k-means와 PSO 알고리즘을 활용한 지능형 시장 세분화 시스템 제안
고객 이탈 예측	단일 모형	오세준, 이은조, 우지영, 김휘강, 2018	온라인 롤플레잉 게임(MMORPG) 사용자 유형을 기반으로 이탈을 예측하는 방법을 제안

	하이브리드 모형	Xie & Li, 2008	Linear discriminant boosting (LD-Boosting)을 활용한 불균형 데이터에서의 고객이탈예측
		Kawale & Srivastava, 2009	게임 플레이어들의 사회적 영향력과 게임에 대한 개인적 참여도를 조사하기 위한 이탈예측모델을 제안
추천 서비스	단일 모형	Tsai & Lu,, 2009	클러스터링 기법과 분류기 앙상블을 결합한 새로운 복합금융손해 모델을 개발 제안
	하이브리드 모형	Xie, Li, Ngai, & Ying, 2009	고객이탈예측을 위한 균형랜덤포레스트(IBRF)방법 제안
		Wen & Zhou, 2012	Dynamic item 클러스터링 기반 협업필터링 추천시스템 제안
		Pham, Cao, Klamma, & Jarke, 2011	소셜네트워크 분석을 통한 협업필터링의 클러스터링 방식 제안
		조윤호, 방정혜, 2011	협업필터링에서의 cold-start 와 데이터 희소성 문제를 해결하기 위한 중심성 분석 제안

제 2 절 고객이탈예측

1) 로열티(Loyalty)

고객이탈의 예측과 방지는 로열티 경영에 있어서 늘 핵심적인 이슈로 연구되어 왔다(김형수, 신동엽, 2012). 로열티(loyalty), 즉 충성도는 고객들의

구매행태에 직접 관련된 행동적 로열티(Behavioral Loyalty)와 구매 외적인 행태와 심리적 애착에 관련된 태도적 로열티(Attitudinal Loyalty)로 구분할 수 있으며, 이 두 가지 로열티의 이차원적 배치에 따라 고객들의 이탈위험도 다르게 나타난다(김형수, 김영걸, 2015). 즉, 행동적 로열티나 태도적 로열티가 각각 높아질수록 고객들의 이탈위험도 역시 낮아지고, 두 가지 로열티가 모두 높은 경우 고객이탈률의 확연한 감소를 보인다. 따라서, 두 가지 특성의 로열티를 모두 고려하여 동질적인 그룹별 이탈관리가 이루어진다면 보다 효율적인 이탈관리가 가능하다(김형수, 김영걸, 2015).

2) 고객이탈예측

한편, 유통업종과 같은 비계약업종은 계약업종이나 구독 비즈니스처럼 명시적인 계약관계가 존재하지 않는 자율거래 업종이기 때문에 고객이탈이라는 명확한 조작적 정의가 존재하지 않으며, 이는 이탈이라는 명확한 목표변수가 존재하지 않으므로 원칙적으로 지도학습 머신러닝 기법으로 해결하기 어렵다(김형수, 신동엽, 2012). 이 경우 고객의 구매패턴을 모델링하여 자신의 구매패턴에서 일정 수준 벗어난 지점을 이탈이라고 간주하는 상대적 이탈예측 모형을 개발해야 한다(김형수, 신동엽, 2012; 오세준, 이은조, 우지영, 김휘강, 2018).

고객이탈예측 연구는 연구의 목적 관점에서 고객이탈을 정의하는 모형을 제시하거나(김형수, 신동엽, 2012), 이탈을 예측하는 모형을 개발하거나(Nath & Behara, 2003; Kraljević & Gotovac, 2010), 고객이탈에 대한 인과관계를 연구하거나(한상린, 맹아름, 2004), 고객이탈 방지를 위한 전략을 제시하는 연구(김충영, 장남식, 김준우, 2003) 등으로 구분할 수 있다. 또한, 연구방법론 관점에서 고객이탈 관련 연구는 가설검증을 통한 실증연구보다는 머신러닝 기법이나 수리적 모델을 활용한 고객이탈 예측 연구가 주를 이루고 있다(Xie & Li, 2008; Kawale & Srivastava, 2009). 이러한 머신러닝 기반의 고객이탈예측 연구는 대부분 이탈예측의 정확도를 높이고자 하는 공통적인 목적을 가지고 있으나, 크게 두 가지의 연구 방향으로 구분된다. 첫

번째 연구 방향은 다양한 머신러닝 알고리즘을 결합한 하이브리드 모델(또는 앙상블 모델)을 활용하여 고성능의 이탈예측 모형을 구축하려는 것이고(오세준, 이은조, 우지영, 김휘강, 2018), 두 번째 연구 방향은 분석에 사용되는 데이터의 불균형 문제를 개선하거나, 예측성능에 영향을 미치는 아웃라이어 (Outlier)를 효과적으로 제거하는 등 모형 개발 이외의 프로세스적인 차원의 개선을 이루려고 하는 연구이다(Tsai & Lu 2009; Xie, Li, Ngai, Ying, 2009).

즉, 머신러닝 기반의 고객이탈예측 연구들은 대부분 예측모형의 개발이나 데이터 전처리와 같은 기술적이고, 지엽적인 문제에 집중되어 온 것이 사실이다. 이러한 기술적인 연구 역시 점차 인공지능화 되는 고객경영의 발전에 반드시 필요한 노력이지만, Python이나 R과 같은 오픈소스를 통해 우수한 머신러닝 기법들을 누구나 사용할 수 있는 상황에서 기술 중심적인 연구의 성과는 자칫 분석 패키지나 함수의 기능에 의존적일 수 있다는 한계가 존재한다. 따라서, 기술적인 모형개발과 더불어 모형을 개발하는 전반적인 프로세스를 전략적인 관점에서 개선하려는 노력 역시 필요한 시점이다.

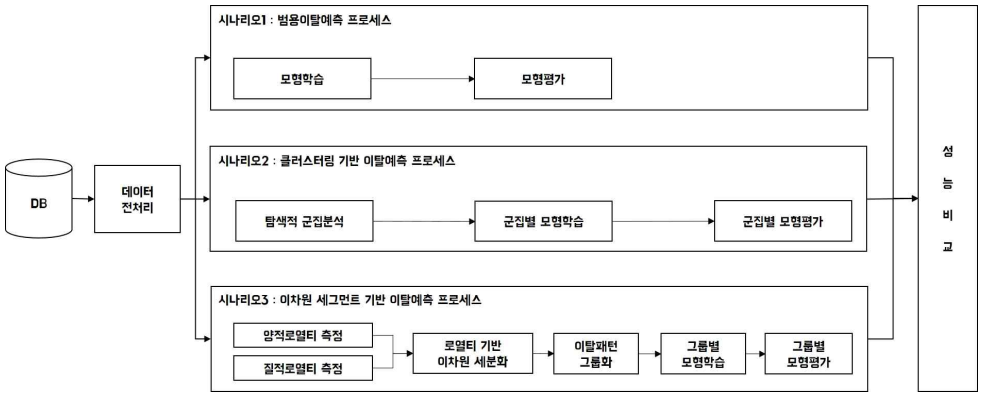
제 3 장 실험 및 평가

제 1 절 모형의 개요

본 연구는 문헌연구 파트에서 기술한 기존 이탈예측 연구의 한계를 극복하고자 이차원 고객충성도 세그먼트 기반의 고객이탈예측 프로세스(CCP/2DL: Customer Churn Prediction based on Two-Dimensional Loyalty segmentation)를 제안하고자 한다. CCP/2DL 프로세스는 고객의 양적, 질적 로열티를 모델링하여 2차원 고객세분화를 수행한 후 도출된 여러 고객 세그먼트를 고객이탈률에 따라 소수의 그룹으로 다시 그룹핑하고, 각 그룹별 최적의 이탈예측 모형을 개별적으로 적용하여 전체 고객에 대한 이탈예측 최적화를 이루고자 하는 방법론이다. 제안 모형은 일반적인 범용이탈예측 프로세스, 클러스터링 기반 이탈예측 프로세스와 함께 성능 비교를 통해 제안모형의 타당성을 입증하고자 하였으며 각각 Scenario 1, 2, 3으로 구분하여 진행하였다. 3가지 이탈예측 프로세스에 사용된 머신러닝 알고리즘은 의사결정나무, 로지스틱 회귀, 랜덤포레스트, 그래디언트 부스팅, 에다부스트, 스태킹 총 6가지이다.

〈그림 3-1〉은 이러한 연구의 전반적인 흐름은 보여주고 있다. Scenario 1은 전체 고객을 하나의 집단으로 간주하고, 학습 데이터 세트와 테스트 데이터 세트로 구분하여 모형을 학습하고 평가하는 범용이탈예측 프로세스이다(Hung, Yen, & Wang, 2006). 이 시나리오 상에서 최적의 모형은 앞서 기술한 6가지 머신러닝 알고리즘을 모두 적용하여 가장 우수한 성능을 보이는 모형을 선택한다. Scenario 2는 의미 있는 고객행태 변수들을 기반으로 탐색적 군집분석을 먼저 시행하여 적절한 군집 수로 고객을 세분화하고, 각 군집 별로 6개의 예측 알고리즘을 적용하여 최적의 모형을 선택하는 클러스터링 기반 이탈예측 프로세스이다(오세준, 이은조, 우지영, 김휘강, 2018). 끝으로 Scenario 3는 앞서 설명한 CCP/2DL 프로세스를 의미하며, 다른 예측 프로세스와 마찬가지로 6가지 예측 알고리즘을 적용하여 각 이탈패턴 그룹별 최적의 모형을 선택하였다. 이탈예측 성능을 평가하기 위해 정확도

(Accuracy), 정밀도(Precision), 재현율(Recall), 특이도(Specificity)와 더불어 정밀도와 재현율의 조화평균으로 계산하는 F1-score를 함께 검토하였다.



〈그림 3-1〉 연구 모델

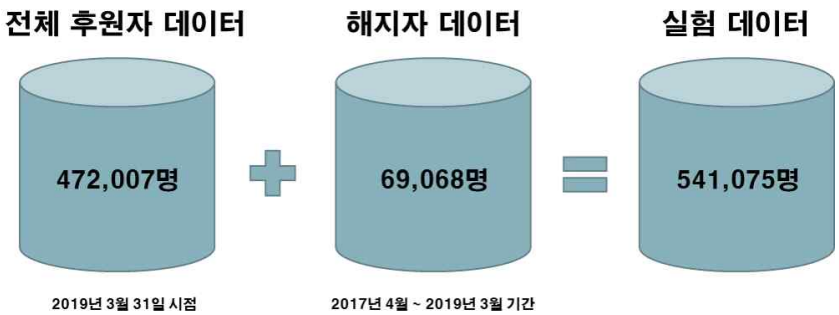
제 2 절 데이터 및 실험설계

1) 실험 데이터

본 연구를 위해 글로벌 NGO 단체 중 하나인 A사의 지원으로 2019년 3월 31일 시점의 472,007명의 전체 후원자와 2017년 4월 ~ 2019년 3월 기간 동안에 발생한 69,068명의 해지자 데이터를 개인 식별정보를 제거하고 사용하였다. 연구에서 사용한 A사 고객 데이터의 전반적인 구성은 〈표 3-1〉과 〈그림 3-2〉와 같다.

〈표 3-1〉 고객 데이터

	Label	Number	Rate	Total
Gender	Male	294,795	54.48%	541,075
	Female	200,248	37.01%	
	Uncertain	46,032	8.51%	
Age	20s	91,499	16.91%	541,075
	30s	98,686	18.24%	
	40s	158,465	29.29%	
	50s	118,428	21.89%	
	60s	53,348	9.86%	
	70s+	20,649	3.82%	
Churn	Not Churned	472,007	87.2%	541,075
	Churned	69,068	12.8%	



〈그림 3-2〉 실험 데이터 구성

고객 데이터에 포함된 200여개의 기본 변수 및 파생 변수 중에서 이탈예측에 유의미하다고 판단되는 122개의 변수를 선별하여 모형 구축에 사용하였다. 이중 이탈예측모형의 속도와 정확도 향상을 위해 특징 선택(Feature Selection)을 사용해 가장 적합한 변수들만을 추출해 모형 구축에 사용하였다. 이때 이탈예측에 사용하기 위해 선별된 122개의 변수는 아래 〈표 3-2〉와 같다.

〈표 3-2〉 데이터 스키마

	변수명	데이터 형태	변수 형태
1	고객ID	식별자	기본변수
2	실제나이	연속형	기본변수
3	성별	범주형	기본변수
4	시/도	범주형	기본변수
5	핸드폰번호 제공 여부	이분형	기본변수
6	이메일 제공 여부	이분형	기본변수
7	우편 수신 여부	이분형	기본변수
8	가입본부	범주형	기본변수
9	총가입기간(일)	연속형	기본변수
10	총가입기간(월)	연속형	기본변수
11	전체 플릿지 수	연속형	기본변수
12	액티브 플릿지 수	연속형	기본변수
13	종료 플릿지 수	연속형	기본변수
14	납부 완료 플릿지 수	연속형	기본변수
15	납부 완료 플릿지 비율	연속형	기본변수
16	액티브 해외아동후원 수	연속형	기본변수
17	액티브 국내아동후원 수	연속형	기본변수
18	개발 채널 수	연속형	기본변수
19	주요 개발 채널	범주형	기본변수
20	개발 미디어 수	연속형	기본변수
21	주요 개발 미디어	범주형	기본변수

	변수명	데이터 형태	변수 형태
22	아동 국가 수	연속형	기본변수
23	아동 최대 나이	연속형	기본변수
24	아동 최소 나이	연속형	기본변수
25	아동 평균 나이	연속형	기본변수
26	액티브 최대후원아동나이	연속형	기본변수
27	액티브 최소후원아동나이	연속형	기본변수
28	정기해외아동 플릿지 수	연속형	기본변수
29	정기국내아동 플릿지 수	연속형	기본변수
30	정기해외사업 플릿지 수	연속형	기본변수
31	정기긴급사업 플릿지 수	연속형	기본변수
32	정기국내사업 플릿지 수	연속형	기본변수
33	정기북한사업 플릿지 수	연속형	기본변수
34	정기전체사업 플릿지 수	연속형	기본변수
35	신규 플릿지 수	연속형	기본변수
36	재후원 여부	이분형	기본변수
37	일시to정기후원 여부	이분형	기본변수
38	피추천유입 수	연속형	기본변수
39	첫플릿지 기준 고객기간	연속형	기본변수
40	납입후원횟수	연속형	기본변수
41	납입총후원금액	연속형	기본변수
42	납입정기후원횟수	연속형	기본변수
43	납입정기후원금액	연속형	기본변수

〈표 3-2〉 계속

	변수명	데이터 형태	변수 형태
44	납입일시후원횟수	연속형	기본변수
45	납입일시후원금액	연속형	기본변수
46	미납 횟수	연속형	기본변수
47	미납율	연속형	기본변수
48	납입일시후원최근성	연속형	기본변수
49	주요납입수단	범주형	기본변수
50	정기해외아동 납입횟수	연속형	기본변수
51	정기국내아동 납입횟수	연속형	기본변수
52	정기해외사업 납입횟수	연속형	기본변수
53	정기긴급사업 납입횟수	연속형	기본변수
54	정기국내사업 납입횟수	연속형	기본변수
55	정기북한사업 납입횟수	연속형	기본변수
56	정기전체사업 납입횟수	연속형	기본변수
57	일시해외선물 납입횟수	연속형	기본변수
58	일시국내선물 납입횟수	연속형	기본변수
59	일시해외사업 납입횟수	연속형	기본변수
60	일시긴급사업 납입횟수	연속형	기본변수
61	일시국내사업 납입횟수	연속형	기본변수
62	일시북한사업 납입횟수	연속형	기본변수
63	일시전체사업 납입횟수	연속형	기본변수
64	정기해외아동 납입금액	연속형	기본변수
65	정기국내아동 납입금액	연속형	기본변수

〈표 3-2〉 계속

	변수명	데이터 형태	변수 형태
66	정기해외사업 납입금액	연속형	기본변수
67	정기긴급사업 납입금액	연속형	기본변수
68	정기국내사업 납입금액	연속형	기본변수
69	정기북한사업 납입금액	연속형	기본변수
70	정기전체사업 납입금액	연속형	기본변수
71	일시해외선물 납입금액	연속형	기본변수
72	일시국내선물 납입금액	연속형	기본변수
73	일시해외사업 납입금액	연속형	기본변수
74	일시긴급사업 납입금액	연속형	기본변수
75	일시국내사업 납입금액	연속형	기본변수
76	일시북한사업 납입금액	연속형	기본변수
77	일시전체사업 납입금액	연속형	기본변수
78	주요 인터랙션 코드	범주형	기본변수
79	주요 문의 유형	범주형	기본변수
80	인터랙션 요청 횟수	연속형	기본변수
81	인터랙션 VOC 횟수	연속형	기본변수
82	인터랙션 캠페인 횟수	연속형	기본변수
83	인터랙션 횟수	연속형	기본변수
84	인터랙션 inbound 횟수	연속형	기본변수
85	인터랙션 outbound 횟수	연속형	기본변수
86	인터랙션 loyalty 횟수	연속형	기본변수
87	인터랙션조건1	연속형	기본변수

〈표 3-2〉 계속

	변수명	데이터 형태	변수 형태
88	인터랙션조건2	연속형	기본변수
89	인터랙션조건3	연속형	기본변수
90	인터랙션조건4	연속형	기본변수
91	종교구분	범주형	기본변수
92	특별후원자구분	범주형	기본변수
93	비전메이커구분	범주형	기본변수
94	메일거부종류수	연속형	기본변수
95	범주형 나이	범주형	파생변수
96	주요후원상품	범주형	파생변수
97	정기해외아동 적합도	연속형	파생변수
98	정기국내아동 적합도	연속형	파생변수
99	정기해외사업 적합도	연속형	파생변수
100	정기긴급사업 적합도	연속형	파생변수
101	정기국내사업 적합도	연속형	파생변수
102	정기북한사업 적합도	연속형	파생변수
103	정기전체사업 적합도	연속형	파생변수
104	일시해외선물 적합도	연속형	파생변수
105	일시국내선물 적합도	연속형	파생변수
106	일시해외사업 적합도	연속형	파생변수
107	일시긴급사업 적합도	연속형	파생변수
108	일시국내사업 적합도	연속형	파생변수
109	일시북한사업 적합도	연속형	파생변수

〈표 3-2〉 계속

	변수명	데이터 형태	변수 형태
110	일시전체사업 적합도	연속형	파생변수
111	상품적합도	연속형	파생변수
112	정보제공 친화성	연속형	파생변수
113	정보수신 친화성	연속형	파생변수
114	수신거부 수	연속형	파생변수
115	납입 안정성	연속형	파생변수
116	일시/선물금 후원 여부	이분형	파생변수
117	일시후원 여부	이분형	파생변수
118	선물금 여부	이분형	파생변수
119	액티브 아동후원 플릿지 수	연속형	파생변수
120	교차후원지수	연속형	파생변수
121	첫후원 소요일	연속형	파생변수
122	이탈여부	이분형	파생변수

전체 데이터의 70%는 모형 학습에, 나머지 30%는 모형 검증에 사용하였다. 데이터 전처리 과정을 통해 범주형 데이터 변환과 이상치 데이터 처리를 시행하였고, 데이터 불균형으로 인한 모델의 성능 감소 문제를 해결하기 위해 오버샘플링 기법을 적용하였다.

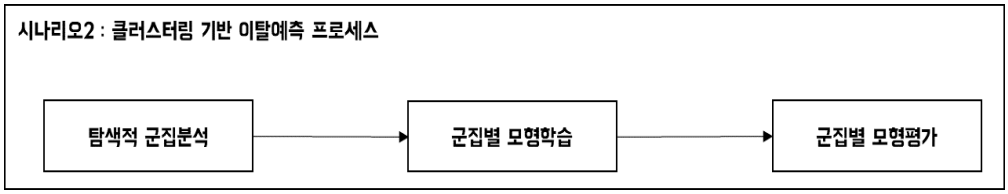
2) 시나리오별 실험설계

첫 번째 시나리오는 범용이탈예측 프로세스이다. 범용이탈예측 프로세스는 고객이탈예측 연구에서 일반적으로 사용되는 방법을 사용하였다. 범용이탈예측 프로세스는 전체 고객을 대상으로 이탈예측 모형을 구축했다는 점이

연구 내 다른 시나리오들과 가장 큰 차이점이다. 본 연구에서는 전체 고객 데이터를 대상으로 앞서 언급한 6가지 이탈예측 알고리즘을 사용하여 모형을 구축하고 성능을 평가하였다.

〈그림 3-3〉 시나리오 1: 범용이탈예측 프로세스

두 번째 시나리오는 클러스터링 기반 이탈예측 프로세스이다. 클러스터링 기반 이탈예측 프로세스는 전처리된 고객 데이터에 탐색적 군집분석을 실행하고 적절한 군집 수로 전체 고객을 군집화한 후 각 군집별로 이탈예측 모형을 구축했다. 본 연구에서 탐색적 군집분석을 실행하기 위해 총 가입기간, 개발 채널 수, 인터렉션 inbound 횟수, 인터렉션 outbound 횟수, 정보친화성 등의 변수를 사용하였으며, 적정 군집의 수는 계층적 군집분석을 실행하고 그 결과에 따라 3개로 결정했다. 분할된 3개의 군집은 앞서 언급한 의사결정나무, 로지스틱 회귀, 랜덤포레스트, 그래디언트 부스팅, 에다부스트, 스태킹 등 총 6개의 이탈예측 알고리즘을 사용해 각각 학습시키고 성능을 비교하여 군집별 성능이 가장 우수한 알고리즘 하나만을 선택했다. 최종적으로 군집당 하나의 최적의 알고리즘으로 구성된 이탈예측모형을 구축하고 성능평가를 진행하였다.



〈그림 3-4〉 시나리오 2: 클러스터링 기반 이탈예측 프로세스

본 연구의 제안 모형인 Scenario 3의 CCP/2DL은 〈그림 3-4〉와 같이 4단계의 프로세스로 구현되었다.

〈그림 3-5〉 시나리오 3: 이차원 세그먼트 기반 이탈예측 프로세스

Step 1: 로열티의 측정

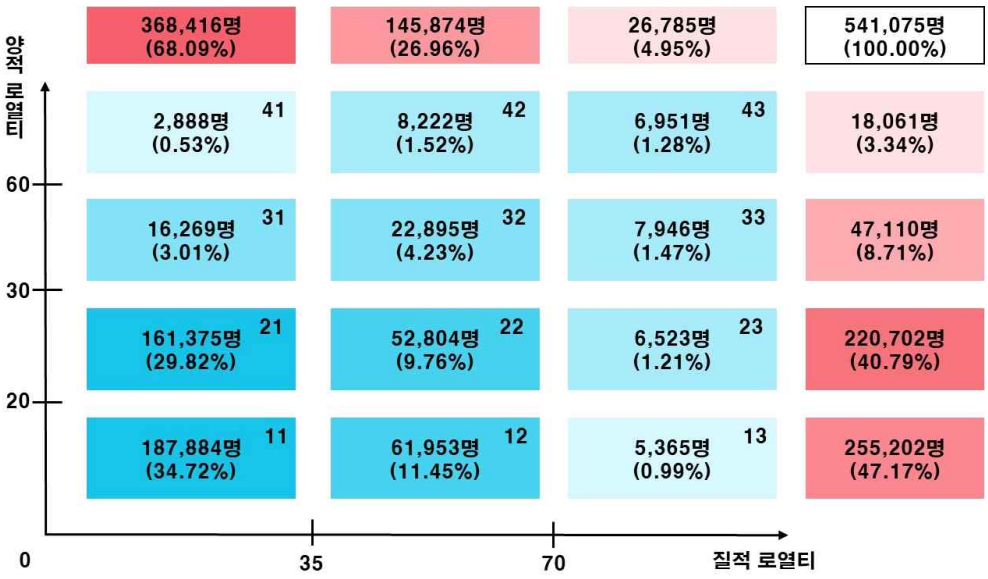
본 연구에서는 고객로열티를 양적 로열티와 질적 로열티로 구분하고, A사와의 협의를 통해 적절한 후보 변수를 선정하여 두 가지 차원의 로열티를 측정하였다. 양적 로열티와 질적 로열티의 측정변수는 〈표 3-3〉과 같이 2개 이상의 고객행태 변수를 요인분석의 요인적재량을 사용하여 구하였다. 고객별 양적 로열티 점수를 종속변수로, 질적 로열티 측정변수들을 독립변수로 하는 다중 회귀분석을 시행하였고, 각 독립변수의 회귀계수를 바탕으로 가중치로 도출하여 결국 후원자별 하나의 질적 로열티 값을 생성하였다.

〈표 3-3〉 이차원 로열티 측정 변수

양적 로열티	총 가입기간, 최근 1년 총 후원금액
질적 로열티	모금채널다양성, 일시후원여부, 정보수신친화성, 인바운드 인터랙션 횟수, 아웃바운드 인터랙션 횟수

Step 2: 로열티 기반 이차원 세분화

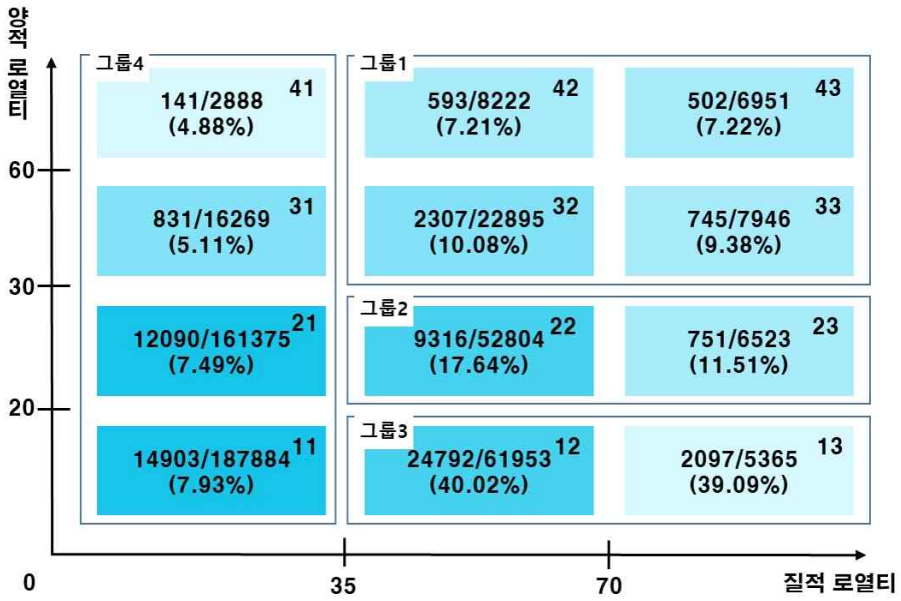
양적/질적 로열티 변수 각각의 분할 기준점을 설정하기 위해 계층적 군집분석과 K-평균 군집화 알고리즘을 사용하였다. K-평균 군집화 알고리즘은 다른 알고리즘들과 비교했을 때, 비교적 대량의 데이터에 대해서 빠르고 안정적인 성능을 보인다(오세준, 이은조, 우지영, 김휘강, 2018). 따라서 상대적으로 용량이 큰 전체 후원자 데이터에 적합하다. 이와 같은 이유로 계층적 군집분석을 통해 파악한 적정 군집 수를 적용하여 전체 데이터를 군집화하는 용도로 사용하였다. 구체적인 방법을 살펴보면 5,000 명을 무작위 추출하여 계층적 군집분석을 수행해 각 로열티 차원의 최적 군집 수를 파악하고, 전체 후원자에 K-평균 군집화 알고리즘을 시행하여 군집 간의 분할 기준점을 도출하였다. 분석 결과, 양적 로열티와 질적 로열티에 대해 각각 4개와 3개의 세그먼트로 분할되어 총 12개의 고객 세그먼트가 생성되었다.



〈그림 3-6〉 Step 2: 로열티 기반 이차원 고객세분화

Step 3: 이탈패턴 분석 및 그룹화

총 12개의 로열티 세그먼트에 대해 이탈패턴으로 재그룹화를 하기 위해 각 세그먼트의 고객 이탈률을 계산하고, 12개의 세그먼트 중 유사한 이탈률을 보이는 세그먼트들을 하나의 그룹으로 병합시켜 총 4개의 고객그룹으로 분류하였다.



〈그림 3-7〉 Step 3: 이탈패턴 분석 및 그룹화

Step 4: 모형 학습 및 평가

이탈패턴 그룹화를 통해 산출된 4개의 그룹에 대해 각각 이탈 예측모형을 개발하고 평가하였다. 모든 그룹에 대해 상기 기술한 6개의 예측모형을 모두 적용하였고, 각 그룹별로 가장 우수한 성능을 보이는 모형을 선택하였다. 최종적으로 선정된 이탈예측 모형에 의해 구간별 이탈위험등급을 설정하고, 실제 이탈률을 재평가한 결과 이탈위험등급이 높을수록 실제 이탈률도 높아져 이탈의 조기경보 시그널로 상당한 효과가 있다는 것을 알 수 있다.

Churn Score	Churn Alerts	Customer		Churned		Churn Ratio in Alert Level
		#	%	#	%	
95~	Very Dangerous	4,481	2.77%	4,406	21.72%	98.33%
80~95	Dangerous	9,727	6.02%	7,957	39.23%	81.80%
50~80	Cautious	18,123	11.21%	4,236	20.88%	23.37%
20~50	Need Attention	51,601	31.92%	2,922	14.41%	5.66%
~20	Safe	77,748	48.09%	763	3.76%	0.98%
Total		161,680	100%	20,284	100%	12.55%

〈그림 3-8〉 Step 4: 이탈위험 등급별 이탈률

제 3 절 경쟁모형의 비교

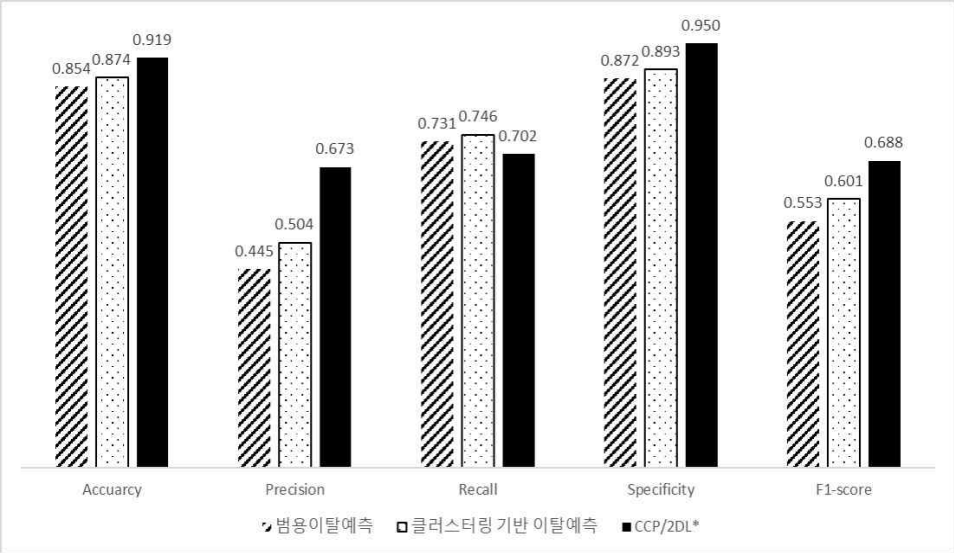
범용이탈예측 프로세스의 최종 선택 모형은 랜덤 포레스트였으며, 클러스터링 기반 이탈예측 프로세스는 2개의 군집에서 랜덤 포레스트가, 나머지 한 개의 군집에서 에다부스트가 선택되었다. 이차원 고객충성도 세그먼트 기반 이탈예측 프로세스의 경우 3개의 고객그룹에서 랜덤 포레스트가, 나머지 한 개 군집에서 그래디언트 부스팅 모델이 선택되었다. 각 시나리오별로 모형의 안정성을 입증하기 위해 데이터 샘플링을 각각 다르게 하여 시나리오별로 총 5번 진행하고 해당 결과의 평균을 시나리오별로 비교하였다. 이러한 분석결과는 〈표 3-3〉에서 제시하고 있다.

각 시나리오별 이탈예측 성과를 평가한 결과 CCP/2DL, 클러스터링 기반 이탈예측, 그리고 범용이탈예측 프로세스의 순으로 예측성과가 뛰어난 것으로 나타났다. CCP/2DL 모형은 5가지 모형평가지표 중 재현율(Recall)을 제외한 4가지 지표에서 다른 이탈예측 방법론에 비해 우수한 성능을 보였으며, 정밀도와 F1-score의 경우 비교적 큰 차이를 보였다. 이러한 분석결과는 〈그림 3-8〉과 〈표 3-4〉에서 제시하고 있다.

성능지표 중 유일하게 낮았던 재현율의 경우 이탈예측모형이 이탈을 엄격하게 판단할수록 다른 지표에 비해 성능이 안 좋을 수 있다. 본 연구에서 사용한 업종은 특성상 후원자들의 기부금이 중단될 경우 기업의 손해가 커지기 때문에 이탈을 예측함에 있어 보수적인 기준을 적용하여 낮은 재현율이 나온 것으로 해석할 수 있다. 또한 재현율과 정밀도의 조화평균을 이용해 계산하는 F1-score가 다른 프로세스들에 비해 우수한 성능을 보여주고 있으므로 상대적으로 낮은 재현율은 문제가 되지 않는 것으로 판단할 수 있다.

〈표 3-4〉 시나리오별 실험 결과

		Accuarcy	Precision	Recall	Specificity	F1-score
1. 범용이탈예측	1회	0.854	0.451	0.701	0.876	0.549
	2회	0.854	0.451	0.701	0.876	0.549
	3회	0.866	0.483	0.798	0.876	0.602
	4회	0.857	0.462	0.748	0.873	0.571
	5회	0.842	0.378	0.708	0.858	0.493
	평균	0.854	0.445	0.731	0.872	0.553
2. 클러스터링 기반 이탈예측	1회	0.879	0.520	0.733	0.900	0.608
	2회	0.868	0.488	0.754	0.885	0.593
	3회	0.867	0.484	0.758	0.883	0.591
	4회	0.875	0.504	0.747	0.894	0.602
	5회	0.880	0.524	0.738	0.901	0.613
	평균	0.874	0.504	0.746	0.893	0.601
3. 이차원 고충성도 세그먼트 기반의 고객이탈예측	1회	0.921	0.678	0.720	0.950	0.698
	2회	0.917	0.667	0.686	0.950	0.676
	3회	0.916	0.667	0.679	0.951	0.673
	4회	0.921	0.679	0.728	0.950	0.702
	5회	0.918	0.676	0.700	0.951	0.687
	평균	0.919	0.673	0.702	0.950	0.688



〈그림 3-9〉 시나리오별 평균성능 비교

〈표 3-5〉 시나리오별 평균성능

	Accuracy	Precision	Recall	Specificity	F1-score
1. 범용이탈예측	0.854	0.445	0.731	0.872	0.553
2. 클러스터링 기반 이탈예측	0.874	0.504	0.746	0.893	0.601
3. CCP/2DL	0.919	0.673	0.702	0.950	0.688

제 4 장 결론 및 향후 연구

제 1 절 연구 요약

본 연구에서는 효과적인 고객이탈 예측을 위해 고객충성도 기반의 이차원 세분화와 이탈패턴 분석을 활용하는 새로운 이탈예측 프로세스 모델을 제안하였다. CCP/2DL이라고 명명한 이 프로세스는 고객의 양적 로열티와 질적 로열티로 기준으로 세분화한 12개의 고객 세그먼트를 도출하고, 유사한 이탈패턴에 따라 다시 4개의 고객그룹을 생성하여 그룹 별 최적의 예측 모형을 적용한다. 제안 모형은 기존의 범용이탈예측 프로세스 및 클러스터링 기반 이탈예측 프로세스와 비교한 결과 상대적으로 우수한 예측성과를 보이는 것으로 나타났다.

제 2 절 연구의 의미

1) 연구의 학술적 의미

본 연구의 학술적 의미로는 첫째, 전체 이탈예측 프로세스의 개선이 예측모형의 개별적인 최적화보다 전체 예측 성능에 더 큰 영향을 미친다는 것을 검증했다는 점이다. 기술적인 관점에서 고객그룹에 적용한 예측 알고리즘과 모형의 학습방식이 동일하기 때문에 본 연구의 실험설계는 프로세스의 차별화를 관찰하고자 하는 방향으로 구성되었다고 볼 수 있다. 둘째, 하나의 전체 고객그룹에 대한 이탈예측 보다는 세분화된 고객그룹에 대한 개별적인 이탈예측이 더 유효하다는 기존 연구의 결과를 재확인하였고, 더 나아가 단순히 기술적인 통계분석에 의한 그룹화보다는 고객로열티와 이탈패턴과 같은 전략적인 기준에 의해 분류된 고객세분화가 이탈예측에 있어 더 효과적이라는 것

을 제시하였다는 점이다. 셋째, 고객 충성도 관련 연구에서 개념적으로 중요시 되어왔던 양적 로열티와 질적 로열티의 개념을 정량적인 측정변수로 모델링하고, 이를 고객세분화에 직접 적용했다는 점이다.

2) 연구의 실무적 의미

본 연구의 실무적 의미로는 첫째, 본 연구의 결과를 사용하여 기업에서는 보다 효율적이고 효과적으로 고객이탈관리를 수행할 수 있다는 점이다. 실제로 본 연구에 참여한 NGO A사는 본 연구결과를 파이썬 모듈로 개발하여 현업에서 후원자 이탈방지에 직접 사용하고 있다. 둘째, 본 연구에서 CCP/2DL과 같은 세분화된 이탈예측 시스템이 운영될 경우 여타 모형에 비해 모형의 유지보수와 전체 최적화에 더 유리하다는 점이다. 전체 고객에 대해 단일 이탈예측모형이 적용될 경우 특정 고객 세그먼트의 행태변화로 인해 전체 이탈예측 모형이 변경되어야 하므로 모형 변경에 따른 리스크가 매우 크다. 셋째, 이탈예측 연구가 상대적으로 일천했던 NGO 업종에 과학적인 CRM 기법을 적용했다는 점이다. NGO의 시장이 점점 커지지만, 경쟁 NGO가 급증하는 현시점에서 NGO 역시 후원자를 '고객(Customer)'으로 인식해야 할 것이며, 이러한 관점에서 NGO에 대한 CRM 전략은 향후 더 중요해질 것으로 판단된다.

제 3 절 연구의 한계점과 향후 연구방향

상기 기술한 연구의 의미와 더불어 본 연구는 몇 가지 연구의 한계점이 존재한다. 우선 본 연구에서 제시한 양적, 질적 로열티 변수는 신뢰성과 타당성 분석이 이루어지지 않았다. 물론 두 가지 로열티 변수의 생성이 가설검증을 위한 개념화나 척도 개발이 아니라, 고객세분화에 사용하기 위한 목적이므로 세분화 변수의 적합성 기준인 동질성(Homogeneity), 이질성(Heterogeneity), 크기(Substantial) (장민석, 김형중, 2018)를 충분히 고려하여

세분화 분류변수를 위한 기술적 타당성은 확보되었다고 볼 수 있다. 그러나, 측정변수와 요인 개발을 위한 전형적인 통계 검증 절차를 거치지 않았으므로 본 연구에서 사용한 두 가지 로열티 변수는 이론적으로 검증된 개념이라고 할 수 없다. 따라서, 향후에는 양적 로열티와 질적 로열티에 대한 이론적 접근과 더불어 신뢰할 수 있는 측정변수로서의 활용을 위해 별개의 신뢰성 및 타당성 검증 연구가 수행되는 것이 바람직하다. 둘째, 본 연구에 적용된 업종은 NGO로서 계약 업종이기 때문에 유통업과 같은 비계약 업종에는 본 연구의 결과를 그대로 활용하기 어렵다는 점이다. 비계약 업종의 경우 계약의 해지와 달리 명시적인 이탈의 개념이 존재하지 않기 때문에 상이한 고객 구매 패턴에 따라 상대적인 이탈기준을 결정하고, 각 이탈수준에 대한 위험도를 측정하는 예측모델을 함수적으로 구현해야 한다. 따라서, 본 연구의 CCP/2DL 프로세스를 적용하되, 본 연구에서처럼 지도학습 머신러닝 모형이 아니라 거래패턴 기반의 이탈예측 모형을 적용하는 연구(김형수, 신동엽, 2012)가 이루어지는 것이 필요하다.

참 고 문 헌

1. 국내문헌

김형수. (2020).『Step by Step 비즈니스 머신러닝 in 파이썬』, 서울: PREDICS

김형수, 김영걸. (2014).『CRM 고객관계관리 전략 : 원리와 응용』, 서울: YOUNG

김형수, 신동엽. (2012). 비계약업종의 이탈관리전략을 위한 상대적 이탈정의 모형 개발.『마케팅연구』, 27(3), 117-144.

김형수, 박예린, 이정민. (2019). 머신러닝을 이용한 공연문화예술 개인화 장르 추천 시스템.『Information systems review』, 21(4), 31-45.

김경태, 이지형. (2018). 딥 러닝과 Boosted Decision Tree를 활용한 고객 이탈 예측 모델.『한국지능시스템학회 논문지』, 28(1), 7-12.

김담희, 안가경. (2018). 머신러닝을 이용한 고객세분화에 관한 연구.『융복합지식학회 논문지』, 6(2), 115-120.

김충영, 장남식, 김준우. (2003). 이동통신서비스 해지고객 예측모형의 비교분석에 관한 연구.『Entrue Journal of Information Technology』, 2(1), 63-74.

박성혁, 김형수. (2012). 고객 이탈방지에 효과적인 멀티 카테고리 상품 추천 방법론.『Entrue Journal of Information Technology』, 11(3), 101-112.

오세준, 이은조, 우지영, 김휘강. (2018). MMORPG 사용자 유형 분류를 통한 이탈 예측 모델 생성 및 평가.『정보과학회 컴퓨팅의 실제 논문지』, 24(5), 220-226.

이건창, 권순재, 신경식. (2001). 은행고객 세분화를 통한 이탈고객 관리분석.『지능정보연구』, 7(1), 177-196.

이재식, 이진천. (2006). 다중모델을 이용한 자동차 보험 고객의 이탈예측.『지능정보연구』, 12(2), 167-183.

장남식. (2005). 사전 세분화를 통한 고객 분류모형의 효과성 제고에 관한 연구.『Information systems review』, 7(2) : 23-40.

장민석, 김형중. (2018). 빅데이터를 활용한 은행권 고객 세분화 기법 연구.『한국디지털콘텐츠학회논문지』, 19(1), 85-91.

정미월. (2008). "의류패션기업의 고객관계관리(CRM)를 위한 RFM고객세분화 모형 설계". 건국대학교 대학원 국내석사학위논문

조영성, 허문행, 류근호. (2008). 모바일 환경하에 RFM 기법을 이용한 개인화된 추천 시스템 개발.『韓國컴퓨터情報學會論文誌』, 13(2), 41-50.

조용민, 남기환. (2017). 협업 필터링 및 하이브리드 필터링을 이용한 동종 브랜드 판매 매장間 취급 SKU 추천 시스템.『지능정보연구』, 23(4), 77-110.

조윤호, 방정혜. (2011). 협업필터링의 신규고객추천 및 희박성 문제 해결을 위한 중심성분석의 활용.『지능정보연구』, 17(3), 99-114.

최도현 박중오. (2015). 빅데이터 환경에서 사용자 거래 성향분석을 위한 머신러닝 응용 기법.『한국정보통신학회 논문지』, 19(10), 2232~2240.

한상린, 맹아름. (2005). 이동통신 상품의 서비스 전환장벽 변화가 고객유지와 고객이탈에 미치는 영향.『商品學硏究』, 23(1) : 89-104.

홍태호, 서보밀. (2004). 인터넷 뱅킹에서 고객의 신념을 이용한 개인화 모형을 위한 데이터마이닝 연구.『인터넷전자상거래연구』, 4(2), 101-115.

2. 국외문헌

Athanassopoulos, A. D. (2000). Customer satisfaction cues to support market segmentation and explain switching behavior. *Journal of business research*, 47(3), 191-207.

Cochocki, A., & Unbehauen, R. (1993). *Neural networks for optimization and signal processing.* (1st), John Wiley & Sons, Inc, New York, NY, USA.

De Bock, K. W., & Van den Poel, D. (2012). Reconciling performance and interpretability in customer churn prediction using ensemble learning based on generalized additive models. *Expert Systems with Applications,* 39(8), 6816-6826.

Duda, R.O., Hart, P.E. and Stork, D.G. (2012), *Pattern Classification.*, New York, John Wiley & Sons.

Elouedi, Z., Mellouli, K., & Smets, P. (2000). Classification with belief decision trees. Aritificial Intelligence: *Methodology, Systems, and Applications(Springer)*, 80-90.

Fathian, M., Hoseinpoor, Y., & Minaei-Bidgoli, B. (2016). Offering a hybrid approach of data mining to predict the customer churn based on bagging and boosting methods. *Kybernetes*.

Hung, C., & Tsai, C. F. (2008). Market segmentation based on hierarchical self-organizing map for markets of multimedia on demand. *Expert systems with applications*, 34(1), 780-787.

Hung, S. Y., Yen, D. C., & Wang, H. Y. (2006). Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3), 515-524.

Kawale, J., Pal, A., & Srivastava, J. (2009). Churn prediction in MMORPGs: A social influence based approach. In 2009 *International Conference on Computational Science and Engineering,* 4, 423-428.

Kraljević, G., & Gotovac, S. (2010). Modeling data mining applications for prediction of prepaid churn in telecommunication services. *Automatika*, 51(3), 275-283.

Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, 38(11), 54-64.

Liao H. Y, Chen K. Y, Liu D. R, and Chiu Y. L. (2015). Customer Churn Prediction in Virtual Worlds. *In Advanced Applied Informatics (IIAI-AAI) 2015 IIAI 4th International Congress on IEEE*, 115-120.

Menard, S. (2002). *Applied logistic regression analysis (Vol. 106)*. USA: Sage.

Nath, S. V., & Behara, R. S. (2003). Customer churn analysis in the wireless industry: A data mining approach. *In Proceedings-annual meeting of the decision sciences institute*, 561, 505-510.

Nie, G., Wang, G., Zhang, P., Tian, Y., & Shi, Y. (2009). Finding the hidden pattern of credit card holder's churn: A case of china. *In International Conference on Computational Science*, 561-569.

Pham, M. C., Cao, Y., Klamma, R., & Jarke, M. (2011). A clustering approach for collaborative filtering recommendation using social network analysis. *J. UCS*, 17(4), 583-604.

Seret, A., Verbraken, T., Versailles, S., & Baesens, B. (2012). A new SOM-based method for profile generation: Theory and application in direct marketin. *European Journal of Operational Research*, 220, 199－209.

Tsai, C.F., & Hung, C. (2014). Modeling credit scoring using neural network ensembles. *Kybernetes*, 43(7), 1114-1123.

Tsai, C. F., & Lu, Y. H. (2009). Customer churn prediction by hybrid neural networks. *Expert Systems with Applications*, 36(10), 12547-12553.

Verdú, S. V., Garcia, M. O., Senabre, C., Marin, A. G., & Franco, F. J.

G. (2006). Classification, filtering, and identification of electrical customer load patterns through the use of self-organizing maps. *IEEE Transactions on Power Systems*, 21(4), 1672-1682.

Wen, J., & Zhou, W. (2012). An improved item-based collaborative filtering algorithm based on clustering method. *Journal of Computational Information Systems*, 8(2), 571-578.

Xie, Y., & Li, X. (2008). Churn prediction with linear discriminant boosting algorithm. *In 2008 International Conference on Machine Learning and Cybernetics,* 1,228-233.

Xie, Y., Li, X., Ngai, E. W. T., & Ying, W. (2009). Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, 36(3), 5445-5449.

Zaïane, O. R. (1999). *Principles of knowledge discovery in databases*. Department of Computing Science, University of Alberta, 20.

ABSTRACT

A Methodology of Customer Churn Prediction based on Two-Dimensional Loyalty Segmentation

Hong, Seung-Woo

Major in Industrial & Management Engineering

Dept. of Industrial & Management Engineering

The Graduate School

Hansung University

Most industries have recently become aware of the importance of customer lifetime value as they are exposed to a competitive environment. As a result, preventing customers from churn is becoming a more important business issue than securing new customers. This is because maintaining churn customers is far more economical than securing new customers, and in fact, the acquisition cost of new customers is known to be five to six times higher than the maintenance cost of churn customers. Also, Companies that effectively prevent customer churn and improve customer retention rates are known to have a positive effect on not only increasing the company's profitability but also improving its brand image by improving customer satisfaction.

Predicting customer churn, which had been conducted as a sub-research area for CRM, has recently become more important as a big data-based performance marketing theme due to the development of

business machine learning technology. Until now, research on customer churn prediction has been carried out actively in such sectors as the mobile telecommunication industry, the financial industry, the distribution industry, and the game industry, which are highly competitive and urgent to manage churn. In addition, These churn prediction studies were focused on improving the performance of the churn prediction model itself, such as simply comparing the performance of various models, exploring features that are effective in forecasting departures, or developing new ensemble techniques, and were limited in terms of practical utilization because most studies considered the entire customer group as a group and developed a predictive model. As such, the main purpose of the existing related research was to improve the performance of the predictive model itself, and there was a relatively lack of research to improve the overall customer churn prediction process. In fact, customers in the business have different behavior characteristics due to heterogeneous transaction patterns, and the resulting churn rate is different, so it is unreasonable to assume the entire customer as a single customer group. Therefore, it is desirable to segment customers according to customer classification criteria, such as loyalty, and to operate an appropriate churn prediction model individually, in order to carry out effective customer churn predictions in heterogeneous industries. Of course, in some studies, there are studies in which customers are subdivided using clustering techniques and applied a churn prediction model for individual customer groups. Although this process of predicting churn can produce better predictions than a single predict model for the entire customer population, there is still room for improvement in that clustering is a mechanical, exploratory grouping technique that calculates distances based on inputs and does not reflect the strategic intent of an entity such as loyalties.

This study proposes a segment-based customer departure prediction process (CCP/2DL: Customer Churn Prediction based on Two-Dimensional Loyalty segmentation) based on two-dimensional customer loyalty, assuming that successful customer churn management can be better done through improvements in the overall process than through the performance of the model itself. CCP/2DL is a series of churn prediction processes that segment two-way, quantitative and qualitative loyalty-based customer, conduct secondary grouping of customer segments according to churn patterns, and then independently apply heterogeneous churn prediction models for each churn pattern group. Performance comparisons were performed with the most commonly applied the General churn prediction process and the Clustering-based churn prediction process to assess the relative excellence of the proposed churn prediction process. The General churn prediction process used in this study refers to the process of predicting a single group of customers simply intended to be predicted as a machine learning model, using the most commonly used churn predicting method. And the Clustering-based churn prediction process is a method of first using clustering techniques to segment customers and implement a churn prediction model for each individual group.

In cooperation with global NGO A, the company conducted analysis and verification using sponsor data, and in conclusion, the proposed CCP/2DL performance showed better performance than other methodologies for predicting churn. This churn prediction process is not only effective in predicting churn, but can also be a strategic basis for obtaining a variety of customer observations and carrying out other related performance marketing activities.

[Keywords] Customer Churn Predictive Model, Customer Relationship Management, Customer Loyalty, Customer Bigdata, CCP/2DL