XAI 를 활용한 통신사 이탈고객의 특성 이해와 마케팅 적용방안 연구

임진희 [1]
[1] 고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
jinheelim@gmail.com

Research on Understanding Churned Customer and Application of Marketing in Telco. industry Using XAI

Jinhee Lim[1]
[1]Dept. of Big Data Convergence, Graduate School of Computer & Information Technology, Korea University

요 약

최근 통신업계에서는 축적된 빅데이터를 활용하여 고객의 특성을 이해하고 맞춤형 마케팅에 이용하려는 노력이 지속되어 왔다. 본 연구에서는 CatBoost 모델을 사용하여 이탈 가능성이 높은 고객을 예측하고 XAI(eXplainable Artificial Intelligence) 기법 중 하나인 SHAP 을 적용하여 이탈에 영향을 미치는 요인을 설명하고자 하였다. SHAP 의 global explanation 기법을 사용하여 특정 고객 segmentation 에 대한 이해력을 높이고, local explanation 기법을 사용하여 개별 고객에 대한 설명과 개인화 마케팅에 적용 가능성을 제시하였다. 본 연구는 기존의 이탈 예측모델인 블랙박스 모델이 갖는 한계점을 극복하고 고객의 특성을 이해하여 실제 비즈니스에 활용 가능성을 높였다는 점에서 의의를 가진다.

1. 서론

통신회사에서는 축적된 빅데이터 자산을 활용한 다양한 연구가 진행되어 왔다. 고객 트래픽 데이터의 특징을 짚어내 불법 행위를 찾아내거나, 고객의 니즈 파악, 고객세분화를 통한 서비스 수요 조사와 반응예측, 이탈 위험이 있는 고객을 미리 선별하여 예방하는 활동 등 다양한 분야에 활용해왔다[1].

그 중에서 고객 이탈에 대한 연구는 오래전부터 진행되어 왔다. 다양한 머신러닝 기법을 적용하여 이탈 고객 예측력 향상[2], 이탈예측과 이를 시각화하여 실제 업무에서의 활용방안 제고[3], 이탈예측에 기반한 고객별 기대수익 예측 등의 연구[4]가 진행되어 왔다. 그러나 이탈 가능성이 높은 고객을 예측하더라도 어떠한 요인들에 의해 이탈이 예상되는지 분석이 되어야 이탈 방어를 위한 구체적인 마케팅 계획 수립이 가능하다.

본 연구에서는 IBM.com 에서 제공하는 미국 캘리포니아 지역의 7,073 명의 고객 데이터[5]를 활용하여 CatBoost 알고리즘으로 이탈 예상 고객을 예측하고, 설명가능한 인공지능 기법을 활용하여 이탈에 영향을 요인을 고객군별, 개별 고객별로 분석하고 이탈방어를 위한 마케팅에 활용 방안을 모색하고자 한다.

2. 이론적 배경

1) CatBoost 알고리즘

CatBoost 는 Categorical Boosting 의 약자로서 범주형 변수(Categorical feature)를 처리하는데 유용한 알고리즘이다. 범주형 변수를 숫자형 데이터로 변환하기 위해 one-hot encoding 등 기존 방식 대신 target statistics 방식을 적용하였다. Target statistic 는 범주형 변수를 같은 범주에 속하는 관측치의 target 값의 통계량으로 대체하는 기법이다[6].

기존의 Gradient Boosting 모델이 동일한 데이터로 계속 학습한 모델로 잔차를 갱신하여 발생하는 Prediction Shift 를 방지하기 위해 CatBoost 에서는 학습 데이터의 일부만 사용하여 모델을 만들고, 이 모델을 이용해 나머지 학습데이터의 잔차를 계산하는 과정을 반복하는 ordered boosting 기법을 사용한다. 또한 트리를 구성하는 매 스텝마다 랜덤으로 섞은 데이터셋을 활용하여 과적합을 방지한다[6].

2) SHAP(Shapley Additive exPlanations)

SHAP 은 각각의 독립 변수에 대한 Shapley value 를 계산하여 독립 변수와 모델의 결과값 사이의 관계를 분석하는 설명 가능한 인공지능 기법이다[7].

SHAP 의 근간이 되는 Shapley Value 는 전체 성과를 창출하는 데 각 특성이 얼마나 공헌했는지를 수치로 표현한다. 각 특성의 기여도는 그 특성의 기여도를 제외하였을 때 전체 성과의 변화 정도로 나타낼 수 있다[8]. Shapley Value 계산 방법은 설명력을 확인하고자 하는 특정변수를 제외한 나머지 변수들의 모든 조합으로 나온 예측값과 특정변수가 있을 경우 예측값과의 차이를 계산하는 것이다. SHAP 은 개별예측결과에 대한 설명력을 제공하는 local interpretation 과 함께 전체 데이터의 변수별 기여도를 시각화하여 설명 가능한 해석을 할 수 있는 global interpretation 을 함께 제공한다[9].

3. 실험평가 및 분석

1) CatBoost 이용한 이탈 예측 모델

본 연구의 통신 데이터는 IBM 에서 제공한 미국 캘리포니아주 7,042 명의 고객 데이터이며, 연령, 성별, 가족 등 기본적인 프로파일과 요금, 사용량, 이용한 서비스 종류, 과금 형태 등을 포함한 32 개 변수로 구성되어 있다. Target 변수는 이탈여부(Churned Value 0//1)이며 그 외 수치형 변수 11 개와, 범주형 변수 20 개로 구성되어 있다. 결측치는 존재하지 않아 결측치 변환을 위한 별도의 전처리는 진행하지 않았다.

전체 데이터셋의 20%를 test set 으로 할당하였고, training set 을 이용하여 예측모델을 학습하였으며, 각 성능 수치는 다음과 같다.

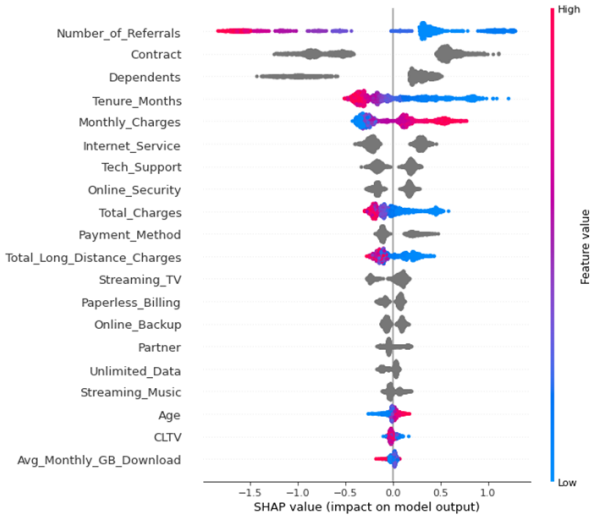
<표 1> CatBoost 알고리즘 Test set 성능

Accuracy	Precision	Recall	F1 Score	AUC-ROC
0.817	0.686	0.588	0.633	0.883

본 연구에 사용된 데이터의 고객이탈율은 26.5%로 이탈 고객보다 유지 고객이 훨씬 많이 포함된 불균형 데이터이다. 불균형 데이터의 경우 소수범주에 속하는 값을 판단하는 것이 중요하므로 Accuracy 대신 F1 score 와 AUC 가 많이 사용된다[10]. 따라서 본 연구에서도 Accuracy 와 함께 F1 Score, AUC-ROC 를 함께 확인하였으며 AUC 는 0.883 으로 양호한 분류 성능을 보여준다.

2) SHAP 을 이용한 고객군 이탈요인 분석

블랙박스 모형인 CatBoost 모델로 이탈고객을 예측하는데 영향을 미친 변수로서는 SHAP summary plot 에서 확인이 가능하다. SHAP summary plot 에서 연속형 변수는 값에 따라 SHAP value 의 크고 작음을 바로 확인이 가능하다. 다만 범주형 변수는 CatBoost 모델에서는 별도의 encoding 없이 사용했으므로 SHAP summary plot 에서는 값의 크고 작음에 따른 구분이 어려워 흑백으로 표시되었다.

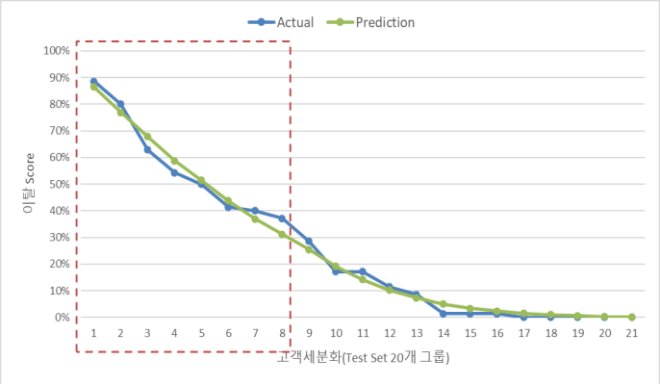


(그림 1) SHAP summary plot of Training set

지인추천(Number of Referrals)이 적을수록, 유지기간(Tenure Months)이 짧을수록, 월이용료(Monthly Charges)가 높을수록 이탈 가능성을 높이는데 SHAP value 가 높게 나타났으며, 이탈가능성에 큰 영향을 미친 것으로 해석된다. 범주형 변수인 계약유형(Contracts)은 단기계약(Month to Month), 가족유무(Dependents)는 가족이 없는 경우, 인터넷연결유형(Internet Services)은 DSL 을 사용하는 경우 이탈가능성을 높이는 것으로 판단할 수 있다.

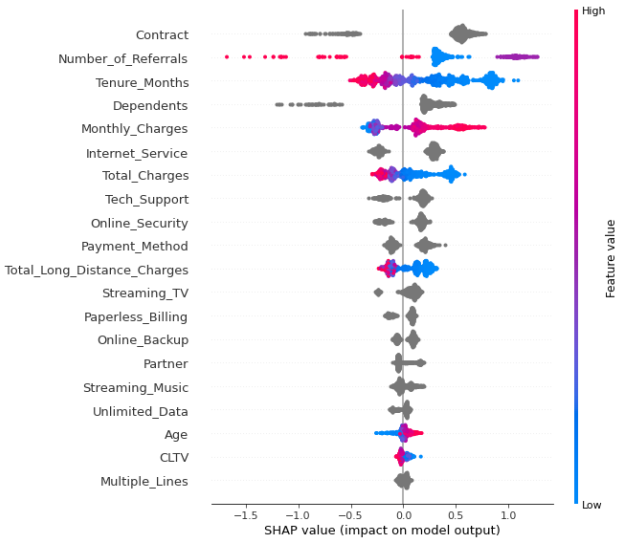
이탈방어 마케팅에 활용 가능성을 확인하기 위해 Test set 데이터에 대해 학습한 CatBoost 모델로 이탈 score 를 산출하고 SHAP 기법을 적용하여 이탈 가능성이 높은 고객의 특성을 확인하고자 했다.

실제 이탈방지를 위한 마케팅 활동에서는 비용, 시간 등의 한계로 모든 고객에게 마케팅 활동을 실행하기 어려울 수 있다. 따라서 이탈 가능성이 높은 상위 일부 고객에 대해서 이탈 방어 활동을 한다는 전제하에 이탈 score 가 높은 일부 고객만 추출하여 이탈가능성을 높이는 요인을 해석하고자 하였다. (그림 2)의 20 개 세분화 그룹 중 이탈 score 가 높은 상위 8 개 그룹에 속하는 고객들을 해석 대상으로 지정했다.



(그림 2) Test Set 이탈예측결과 Lift Chart

이들 고객에 대해 마케팅 활동 전 고객그룹의 특성을 확인하기 위해 SHAP 기법을 적용하였다. 그 결과, 이탈 가능성에 영향을 미치는 상위 변수는 Training set 과 유사하게 나타났으나, 순서로는 Training set 의 지인추천수와는 달리 계약유형이 가장 큰 영향을 미친 것으로 나타났다.



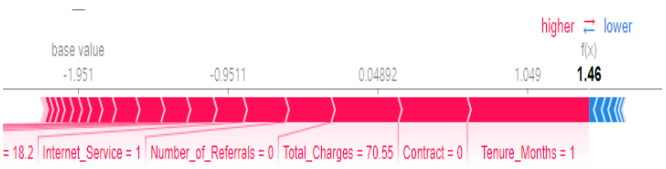
(그림 3) SHAP summary plot of Test set(score 상위 40%)

계약유형(Contracts) 변수의 실제 값인 단기계약(Month-to-Month) 특성상 한달 단위로 계약을 연장해야 하므로 도중에 계약을 연장하지 않을 가능성이 높은 것으로 해석할 수 있다. 또한 총계약기간(Tenure Months)이 짧을수록 이탈가능성이 높은 것으로 나타나 계약초기에 이탈할 가능성이 높은 것으로 해석된다. 따라서 이들 고객에게는 단기계약을 장기계약으로 유도하는 마케팅이 필요하고 계약초기에 유지를 위한 프로모션을 검토해볼 수 있다.

3) SHAP 을 이용한 개별 고객 이탈요인 분석

SHAP force plot 은 개별 인스턴스(고객)에 대한 SHAP value 와 변수별 영향력을 제공해준다.

(그림 4)는 이탈 score 가 79.9%인 고객에 대해 이탈가능성을 높이는데 영향을 미친 변수를 제시하고 있다. 이 고객은 계약유지기간이 한달, 계약형태가 Month to Month, 총 이용료가 70.55 인 고객으로 가입 후 한달 이용 후 해지가 예상되는 고객이다. 한달 동안의 이용료가 70.55 로 다른 고객의 월 평균 이용료 65.71 보다 다소 높은 편으로 나타났다.



(그림 4) 이탈 score 가 79.9%인 고객의 SHAP value

이 고객에게는 이용료 할인을 받을 수 있는 할인프로그램 제안과 함께 월단위 단기 계약이 아닌 장기계약으로 유도하는 맞춤형 마케팅 활동을 고려해 볼 수 있다.

4. 결론 및 향후 과제

본 연구에서는 범주형 변수가 많은 데이터 특성에 적합한 Catboost 알고리즘을 활용하여 이탈 고객을 예측하였다. 그리고 통신사 고객 이탈 방어를 목적으로 이탈 고객의 특성 이해와 마케팅 방안에 연계를 위해 SHAP 이라는 XAI 기법을 이용하여 이탈에 영향을 미치는 특성변수를 확인하여 이탈고객의 특성을 설명하고자 하였다.

SHAP 에서 제공하는 global explanation 기법을 활용하여 세분화된 고객군별 이탈가능성과 함께 이탈에 영향을 미치는 요인을 설명하여 적절한 마케팅방안 수립에 참고할 수 있다. 또한 SHAP 에서 제공하는 local explanation 기법을 활용하여 개인별 이탈에 미치는 요인을 설명하여 개인별 차별화된 대응 전략 수립에 참고할 수 있다.

다만 본 연구에서 활용한 통신사 고객 데이터는 고객의 프로필 정보, 계약 관련 정보 등으로 한정되어 있다. 최근 통신업계 상황이나 최신 상품 이용 정보가 반영되지 않았고 접속 정보, 트래픽 정보 등 행동정보가 반영되지 않아 최근 통신사 이용고객의 행동특성을 설명하기에는 한계가 있다. 따라서 향후 연구에서는 최신 데이터를 이용하여 이탈 예측의 정확도를 높이고 설명력을 높일 수 있는 연구가 필요하다.

참고문헌

[1] aimap marker, “통신업계의 10 가지 대표적인 데이터 과학 활용 사례”, https:// medium.com/@aimap.marker /통신업계의-10 가지-대표적인-데이터-과학-활용-사례-c1b2650b4685

[2] Park, Bo Ram, “A Study on the Application of Datamining in predicting customer churn of the mobile phone company”, Master’s thesis, Ewha Womans University, 2011

[3] S.-H. Kim, K.-W. Kim, Y.-S. Kim, T.-Y. Yoon, and J.-W. Jeon, “Analysis of customer churn prediction in telecom industry using Machine learning & Deep learning,” Proceedings of the Korea Information Processing Society Conference, pp. 568–571, Nov. 2020.

[4] Jaeyeop Kim, “Profit based model selection and optimization for customer retention”, Master’s thesis, Seoul National University, 2020

[5] IBM Accelerator Catalog, https://community.ibm.com /accelerators/catalog/content/Telco-customer-churn-status-and-reason-for-leaving, 2020

[6] L. Prokhorenkova, G. Gusev, A. Vorobev, A. Dorogush, A. Gullin, “CatBoost: unbiased boosting with categorical features”, Moscow Institute of Physics and Technology, 2017

[7] S. Park, Y. Noh, S. Jung, and E. Hwang, “SHAP-based Explainable Photovoltaic Power Forecasting Scheme Using LSTM,” Proceedings of the Korea Information Processing Society Conference, pp. 845–848, Nov. 2021

[8] Seonghun Kim, Woojin Kim, Yeonju Jang, Hyeoncheol Kim.(2021).Development of Explainable AI-Based Learning Support System.The Journal of Korean Association of Computer Education,24(1),107-115.

[9] D. Kim, A. Jeong, and T. Lee, “Analysis of Malware Group Classification with eXplainable Artificial Intelligence,” Journal of the Korea Institute of Information Security & Cryptology, vol. 31, no. 4, pp. 559–571, Aug. 2021.

[10] B.-H. Lee, T.-H. Kim, and Y.-S. Choi, “Discriminant analysis for unbalanced data using HDBSCAN,” The Korean Journal of Applied Statistics, vol. 34, no. 4, pp. 599–609, Aug. 2021.