SHAP-LLM 기반 IPTV 가입 예측 및 개인화된 Sales Talk 생성 프레임워크 - SK브로드밴드 사례

백은진*, 김민정

*서경대학교, SK브로드밴드

*eunjin@skuniv.ac.kr, minjung.kim@sk.com

A SHAP-LLM-Based Framework for IPTV Subscription Prediction and Personalized Sales Talk Generation - A Case Study of SK Broadband

Eunjin Baek*, Minjung Kim

*Seokyeong University, SK Broadband

요 약

본 연구는 SK브로드밴드 고객 중 인터넷 서비스만 이용하는 고객을 대상으로 IPTV 가입 가능성을 예측하는 머신러닝 모델을 개발하였다. 총 25개의 특성 변수를 활용하고, 불균형 클래스 문제는 클래스 가중치 조정, 하이퍼파라미터 최적화, 교차 검증 기법 등을 통해 보완하였다. 모델의 예측값이 높은 고객을 중심으로 SHAP(SHapley Additive exPlanations) 기법을 활용해 예측에 영향을 준 주요 변수를 도출하였다. 이 변수들은 LLM(GPT) 입력값으로 구조화되어, IPTV 가입 유도를 위한 개인화된 Sales Talk를 자동 생성하는 데 활용하였다. 본 연구는 단순 예측을 넘어, 예측-해석-생성을 연계한 자동화 프레임워크를 설계하여 실제 마케팅 활용 가능성을 제시하였다.

Ⅰ. 서 론

통신 시장의 결합 상품 경쟁이 심화되면서, 고객 유치와 매출 증대를 위한 정교한 마케팅 전략의 필요성이 커지고 있다. 특히 IPTV는 ARPU(고객당 평균 매출) 향상과 고객 이탈 방지에 기여함에 따라, 인터넷 단독 이용 고객을 위한 가입 유도 전략이 중요해지고 있다. 하지만 기존 마케팅은 고객 개별 특성을 반영하지 못하고, 인구 통계 중심 전략에 머무르는 한계를 보였다. 선행 연구에 따르면, IPTV 가입에 영향을 미치는 요인의 효과는 고객 특성에 따라 달라지는 조절 효과로 확인되었으며, 이를 통해 고객 특성 기반 전략의 중요성이 강조되었다[1]. 기존 연구는 IPTV 가입 요인을 설명하는 데 초점을 맞추었지만, 본 연구는 이를 확장하여 SK브로드밴드 고객을 대상으로 예측-해석-생성 구조를 적용한 통합 프레임워크를 제안한다. 제안된 프레임워크는 IPTV 가입 가능성이 높은 고객을 사전 식별하고, 개별 특성에 기반한 Sales Talk를 자동 생성하는 전 과정을 하나의 흐름으로 연결하였으며, 실제 적용 가능성을 고려해 구현까지 수행되었다.

Ⅱ. 본 론

본 연구의 전체 분석 절차는 그림 1과 같이 총 4단계로 구성된다. 고객 예측부터 Sales Talk 생성까지의 전 과정을 하나의 흐름으로 연결한 점이 특징이다.

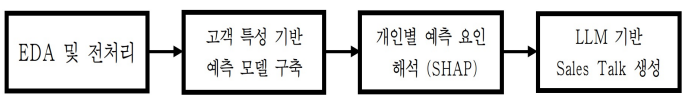

그림 1. 예측-해석-생성 기반 프레임워크

1. 데이터 EDA 및 전처리

본 연구는 2025년 특정 시점을 기준으로 확보된 SK브로드밴드 실제 고객 데이터 50,000건을 활용하였으며, IPTV 가입 여부(iptv_yn)를 예측하는 이진 분류 문제로 설정하였다. 전체 고객 중 IPTV 가입자는 약 89.16%, 미가입자는 약 10.84%로, 클래스 불균형이 존재하였다. 독립변수 중 타겟 변수와 직접적으로 연관된 항목은 분석에서 제외하였으며, 변수 간 상관계수를 분석하여 다중공선성이 우려되는 항목을 제거하였다. 또한 EDA 결과를 바탕으로 예측 구분력이 낮은 변수도 제외하였다. 이상치는 수치형 변수의 상위 1% 극단값을 제외하였고, 결측치는 변수 특성에 따라 최빈값 또는 중앙값으로 대체하였다. 범주형 변수는 LabelEncoder, 연속형 변수는 StandardScaler로 전처리하였다.

2. 모델링 및 성능 비교

전처리된 데이터셋을 바탕으로 IPTV 가입 여부(iptv_yn)를 예측하기 위해 XGBoostClassifier, LightGBMClassifier, RandomForestClassifier, CatBoostClassifier의 네 가지 트리 기반 분류 알고리즘으로 실험하였다. 각 모델에 대해 Optuna를 활용하여 하이퍼파라미터 튜닝을 수행하였으며, Trial 수를 점진적으로 증가시킨 후 최고 성능 인근에서 국소 탐색(Local Search)을 적용하였다. XGBoostClassifier는 학습률(learning_rate), 최대 깊이(max_depth), 반복 횟수(n_estimators) 등 핵심 파라미터를 중심으로 최적화되었다. 모델 튜닝 이후, 성능 평가는 Stratified K-Fold 교차 검증(5-Fold)을 기반으로 수행하였다. 클래스 불균형 문제를 고려하여 Precision과 Recall의 균형을 종합적으로 평가할 수 있는 F1 Score(Weighted Average)를 주요 지표로 선정하였으며, ROC-AUC도 함께 활용하였다. 불균형 문제는 class_weight='balanced'와 Stratified Split 방식으로 대응하였다[2]. 또한 SelectFromModel 기법을 적용해 중요도가 낮은 변수를 제거하는 실험도 수행하였다. 그러나 성능 개선 효과는 나타나지 않아, 최종 모델에서는 전체 변수를 유지하였다. 모델 비교 결과, XGBoostClassifier가 F1 Score(0.8865)와 ROC-AUC(0.9032) 모두에서 가장 우수한 성능을 보여 최종 예측 모델로 선정하였다. 본 연구의 모델 비교 결과는 표 1에서 확인할 수 있다.

모 델	F1 Score	ROC-AUC	Precision	Recall
XGBoost Classifier	0.8865	0.9032	0.8844	0.8890
LightGBM Classifier	0.8470	0.8904	0.8904	0.8264
RandomForest Classifier	0.8500	0.8960	0.8900	0.8270
CatBoost Classifier	0.8364	0.8927	0.8926	0.8103

표 1. 분류 모델별 성능 비교 결과

3. SHAP 기반 예측 요인 해석

최종 선정된 XGBoostClassifier 모델의 예측 결과 해석을 위해 SHAP(SHapley Additive exPlanations) 값을 활용하였다[3]. SHAP은 게임 이론에 기반한 기여도 설명 기법으로, 개별 고객 예측에 영향을 준 주요 요인을 정량적으로 파악할 수 있다. 분석 대상은 실제 IPTV 미가입자 중에서도, XGBoostClassifier 모델이 예측 확률 0.5 이상으로 분류한 가입 예상 고객군이다. 이들은 마케팅 타겟 고객으로 간주되며, 영향 요인을 도출하기 위해 SHAP 분석을 수행하였다. 가장 자주 도출된 상위 기여 변수는 '인터넷 상품명', 'SK 와이파이 공유기 사용 여부', '월별 와이파이 접속 기기 수'였다. 그림 2는 XGBoostClassifier가 예측 확률 0.792로 분류한 IPTV 미가입 고객 사례의 SHAP Waterfall Plot이다. 모델 기준값(base value)인 0.5에서 시작해, 각 변수의 영향력이 누적되며 최종 예측값이 형성되는 구조이다. 빨간색 막대는 예측값을 높이는 양의 기여 요인을, 파란색 막대는 예측값을 낮추는 음의 기여 요인을 의미한다. 해당 고객의 경우, '인터넷 상품명'이 예측값을 +1.99만큼 증가시켜 가장 큰 양의 영향을 미쳤으며, '연령대', '방문 기사 권유 구매 여부', '인터넷 회선 수' 등은 음의 방향으로 작용하였다. 이러한 시각화는 고객별 예측 요인을 직관적으로 해석할 수 있게 해주며, 이후 LLM 기반 Sales Talk 생성의 핵심 입력으로 활용되었다. 이처럼 SHAP Waterfall Plot은 모델의 의사결정 과정을 수치적으로 해석하는 데 유용하다[4].

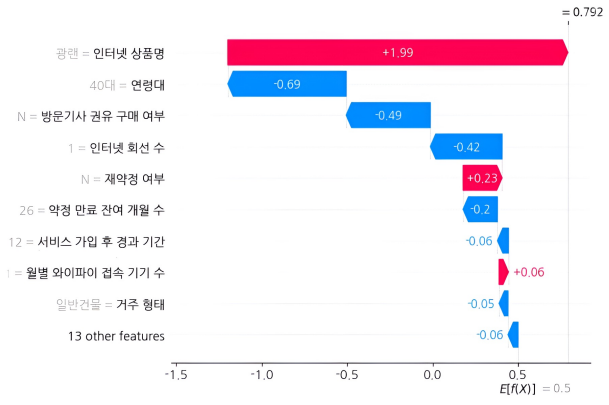

그림 2. SHAP Waterfall Plot: 개별 고객 예측 기여 요인 시각화

4. SHAP 기반 주요 요인의 구조화 및 LLM 프롬프트 설계

고객별 SHAP 분석 결과로 도출한 상위 3개 주요 특성을 기반으로 Sales Talk 생성을 위한 입력값을 구성하고, 이를 통합한 구조화 파일을 LLM 입력 데이터로 활용하였다. LLM 입력값은 SHAP 25개 변수 중 성별·연령대 등 의미 전달이 모호하거나 Sales Talk 생성을 위한 표현이 어려운 16개 항목을 제외하고, 9개 변수만 활용하였다. 프롬프트는 LLM이 상담사처럼 자연스럽게 응답하도록 설계되었으며, 변수명은 직관적인 표현으로 변환하였다. 예를 들어, 'WingsCount'는 '와이파이 확장기 개수', 'Monthly Device Number'는 '월별 와이파이에 연결된 기기 수'로 변경했다. 또한 숫자형 변수는 단순 수치 대신 범주화하였으며, Sales Talk 생성 시 의미가 반영될 수 있도록 구성하였다. 예를 들어, MonthlyDeviceNumber 값이 3인 경우는 사전 기준(2 이상)에 따라 '와이파이에 여러 기기를 연결해 사용하는 고객'으로 해석되어 Sales Talk에 반영되었다. 와이파이에 연결된 기기 수는 고객의 실제 사용 기기 수를 유추할 수 있는 지표로 간주하였다. 프롬프트는 few-shot 방식으로 설계되어, 예시를 함께 제공함으로써 LLM이 상담 문장의 형식과 문체를 학습할 수 있도록 적용하였다[5]. 표 2는 SHAP 분석을 통해 도출된 한 고객의 주요 변수와 실제 값 및 생성된 Sales Talk 예시를 보여준다.

고객 ID	SHAP 주요 변수 및 실제값	생성된 Sales Talk 예시
23	(1) 인터넷 상품명 = 기가라이트 (2) 월별 와이파이 접속 기기 수 = 6 (3) SK 와이파이 공유기 사용 여부 = Y	여러 기기를 와이파이에 연결해 이용 중인 고객님께는, 기가라이트 인터넷과 SK 와이파이 공유기를 통해 IPTV 콘텐츠를 더욱 안정적이고 원활하게 즐기실 수 있습니다.

표 2. SHAP 주요 변수 및 실제 값 기반 LLM Sales Talk 생성 예시

Ⅲ. 결 론

본 연구는 IPTV 마케팅 자동화를 위해 예측-해석-생성의 전 과정을 통합한 프레임워크를 제안하였다. 단순한 가입 예측을 넘어 개별 고객의 예측 근거를 SHAP 기반으로 해석하고, 이를 LLM을 활용한 Sales Talk로 전환함으로써 실무 적용 가능성을 제시하였다. XGBoostClassifier 기반 예측 모델은 불균형 데이터를 효과적으로 처리하며 네 가지 모델 중 가장 우수한 성능을 보였으며, SHAP 분석을 통해 IPTV 가입 유무에 영향을 미친 고객별 주요 요인을 해석할 수 있었다. 주요 변수는 의미 매핑을 거쳐 자연어 문장으로 변환되었으며, LLM을 활용해 IPTV 가입 유도를 위한 개인화된 Sales Talk로 전환되었다. 예측-해석-생성 구조는 Sales Talk 자동화를 통해 고객 대응의 일관성을 높이고, 맞춤형 전략 수립에 기여할 수 있다. 향후 연구에서는 생성 멘트의 품질과 설득력을 검증해 실무 적용성을 강화할 수 있을 것이다. 또한, 다양한 프롬프트 설계나 오버샘플링 기법에 대한 후속 실험도 향후 보완될 수 있다.

참 고 문 헌

[1] 배형우, 김도형, 『IPTV 서비스 가입 의도에 대한 연령대 조절 효과 분석』, 『방송공학회논문지』, 제29권 제1호, 2024, 쪽 81-93.

[2] Q. Gao, et al., "Comprehensive review of class imbalance learning techniques," *Journal of Machine Learning Research*, vol. 26, no. 112, pp. 1-33, 2025.

[3] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[4] T. Saito and S. Nakagawa, "Practical guide to SHAP analysis: Explaining supervised machine learning models," *Frontiers in Pharmacology*, vol. 15, 11513550, 2024.

[5] Y. Zhu, Y. Zhang, J. Liu, and W. Yang, "Generative AI for marketing: A new frontier," *Journal of Interactive Marketing*, vol. 61, pp. 25-39, 2023.