머신러닝 기반 VOD 월정액 서비스 가입 추천 고객 예측 및 LLM 활용 개인화된 Sales Talk 생성

*하담, 김민정
*건국대학교, SK 브로드밴드

*andy@konkuk.ac.kr, minjung.kim@sk.com

Predicting VOD Subscription Target Customers Using Machine Learning and Generating Personalized Sales Talks with LLMs

*Ha Dam, Kim Minjung
*Konkuk University, SK Broadband

요 약

본 연구는 통신사의 VOD 월정액 서비스 가입률 향상을 위해 SK브로드밴드의 약 5만 명 고객 데이터를 활용하여 월정액 서비스 미가입자 중 가입 가능성이 높은 고객을 선별하는 머신러닝 기반 분류 모델을 개발하였다. 고객의 인구통계 정보와 인터넷·IPTV 사용 패턴 등 43개 변수로 모델을 구축하고, 데이터 불균형 문제를 해결하기 위해 클래스 가중치 조정과 계층적 교차검증(stratified cross-validation) 기법을 적용하였다. 하이퍼파라미터 최적화와 특성 선택(feature selection)을 통해 최적의 모델을 선정하고 미가입자를 대상으로 가입 여부를 예측하였다. 이후 예측된 추천 고객군을 대상으로 SHAP 분석을 수행하여 가입 예측에 기여한 주요 요인을 식별하고, 이를 대규모 언어 모델(Large Language Model, LLM)과 연계하여 고객별 맞춤형 가입 권유 문구(Sales Talk)를 자동 생성하는 프레임워크를 제안하였다.

Ⅰ. 서 론

통신사의 주요 부가서비스 중 하나인 VOD 월정액 서비스는 고정 요금을 내면 VOD 콘텐츠를 한 달간 무제한으로 시청할 수 있는 구독형 서비스이다. 방송통신위원회에 따르면, 국내 OTT 시장 규모는 2023년 5.6조 원에서 2027년에는 7.2조 원까지 성장할 것으로 전망된다. 미디어 소비 패턴이 실시간 채널 중심에서 VOD 중심으로 전환되면서, 구독형 VOD 서비스는 통신 산업 내 전략적 중요성이 증대되고 있다.

기존 연구에서는 고객의 서비스 가입 여부를 분류모델로 예측하고, SHAP 분석을 통해 모델의 의사결정 과정을 설명한 바 있다[1]. 최근에는 대규모 언어 모델(LLM)을 활용하여 SHAP 분석 결과를 연계함으로써 모델의 의사결정 과정을 직관적으로 해석할 수 있도록 하는 방법이 제안되었다[2]. 본 연구는 통신사의 VOD 월정액 서비스에 대해, 머신러닝 기반 이진 분류 모델을 활용하여 미가입 고객 중 가입 가능성이 높은 추천 대상을 예측하고, 고객별 SHAP 분석 결과를 LLM과 연계하여 개인화된 가입 권유 문구(Sales Talk)를 자동 생성하는 프레임워크를 제안한다. 본 연구의 차별점은 예측 모델의 해석 결과를 개인화 추천 문구 생성으로 연결하여 실제 마케팅 액션화함으로써, 데이터 기반 마케팅의 실질적 효과를 높이는 데 있다. 이를 통해 예측 모델의 설명력을 마케팅 실무에 직접 활용함으로써, 서비스 가입률 향상과 마케팅 효율성 제고에 기여할 수 있을 것으로 기대한다.

Ⅱ. 본론

1. 데이터 EDA 및 전처리

본 연구에서 모델 학습과 평가에 사용된 데이터는 2025년 2월 기준 SK브로드밴드의 고객 중 층화 표본추출(stratified sampling)을 통해 수집된 약 5만 명의 데이터이다. VOD 월정액 서비스 가입자와 미가입자가 함께 있는 데이터로 고객의 독립변수들에 따라 월정액 서비스 가입 여부를 학습시켜 모델이 가입자와 미가입자를 분류하도록 하였다. 독립변수(X)는 총 43개로, 인구통계학적 특성(연령대, 성별, 가입 지역 등), 서비스 가입 정보(인터넷 및 IPTV 상품명, 셋톱박스 모델명, 서비스 가입 기간 등), 그리고 미디어 소비 행태(실시간 채널 및 VOD 시청 시간 등)로 구성되어 있다. 종속변수(Y)는 'PPM_YN'으로, VOD 월정액 서비스의 가입 여부를 나타낸다. 전체 고객 중 가입자(Y)와 미가입자(N)의 비율은 약 1:8.1로, 심각한 클래스 불균형(class imbalance)이 존재하였다.

독립변수들 중 월정액 서비스 가입 여부에 영향을 미치는 변수들을 탐색한 결과, 서비스(인터넷/IPTV) 가입 기간, 키즈 가구 여부, IPTV와 모바일 페어링 여부, 실시간 채널 시청 시간 등의 값에 따라 유의미한 가입률 차이가 나타났다. 특히 VOD 시청 등급 및 구매 등급과 같은 VOD 관련 변수들에서 가입률의 큰 차이를 보였는데, 이는 월정액 서비스의 이용 이력이 VOD 시청·구매 기록에 포함되기 때문으로 해석된다. 한편, VOD 구매 이력이 없으나 시청 시간이 존재하는 고객군도 확인되었으며, 이는 무료체험 이용 고객임을 확인한 후 ‘ FreeTrial’ 이라는 파생 변수를 추가 생성하였다.

데이터 전처리 과정에서 25개 변수의 공통 결측치를 포함한 637개 행을 제거하고, 나머지는 변수 특성에 따라 수치형은 중앙값 또는 0, 범주형은 ‘ 기타’ 로 처리하였다. 변수 간 상관관계를 분석하여 절댓값 기준 0.7 이상의 높은 상관성을 가진 변수들 중 의미가 중복되는 5개 변수를 제거하였고, 정답의 의미를 내포하고 있는 변수 역시 제외하였다. 또한, 수치형 변수 중 분포 상 극단적으로 큰 이상치가 확인된 변수는 상위 1%를 기준으로 클리핑(clipping) 하여 극단 값의 영향을 완화하였다. 범주형 변수에는 레이블 인코딩(label encoding)을, 수치형 변수에는 표준화(standardization)를 통해 스케일링 처리하였으며, 종속변수(Y)는 이진 분류를 위해 Y와 N을 각각 1과 0으로 변환하였다.

2. 모델링 및 결과

본 연구에서는 불균형 데이터에서도 효과적인 성능을 보이는 트리 기반 앙상블 분류 모델인 XGBoost Classifier, CatBoostClassifier, LGBMClassifier, RandomForestClassifier 총 4개의 모델을 비교 실험하였다. 전처리가 완료된 49,363개의 데이터는 층화추출(stratified sampling)을 통해 학습용과 테스트용 데이터로 4:1 비율로 나누었다. 학습 데이터에 대해서는 Stratified K-Fold 방식의 5-Fold 교차검증을 적용하여 모델의 일반화 성능을 확보하고 클래스 불균형 문제에 대응하였다.

불균형 데이터 문제를 완화하기 위해, XGBoost Classifier, CatBoostClassifier, LGBMClassifier에는 클래스 가중치 파라미터인 scale_pos_weight를 적용하였으며, RandomForestClassifier에는 class_weight ='balanced' 설정을 사용하였다[3].

각 모델의 하이퍼파라미터는 Optuna를 활용하여 교차검증에서 Class 1의 평균 F1-score를 최대화하도록 최적화하였다. 이후 각 모델의 feature importance를 기준으로 상위 k개의 변수를 선택하는 방식으로 Feature Selection을 수행하였으며, k= 30~43 구간에서 교차검증 기반 Class 1의 F1-score 평균값을 비교하여 가장 높은 성능의 k로 최종 feature 개수를 선정하였다. 본 연구는 소수인 양성 클래스의 가입자를 정확히 분류하는 것이 목적이므로 교차검증 기반 평균 Class 1의 F1-score와 PR-AUC를 주요 평가지표로 삼았다.

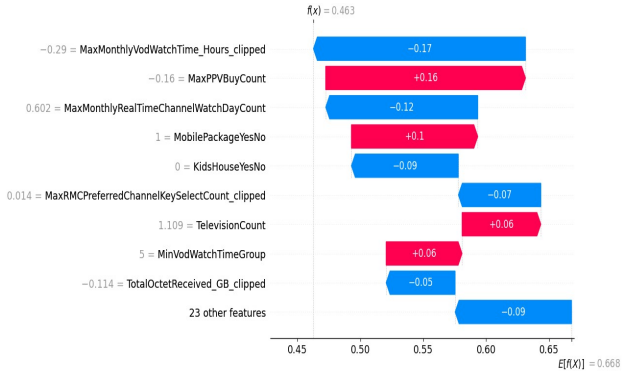
	Class 1 Precision	Class 1 Recall	Class 1 F1	PR-AUC
XGBoostClassifier (k=32)	0.7168	0.7049	0.7107	0.7782
CatBoostClassifier (k=36)	0.6895	0.7275	0.7078	0.7709
LGBMClassifier (k=34)	0.6944	0.7273	0.7104	0.7723
RandomForestClassifier (k=41)	0.6825	0.7206	0.7009	0.7563

[표 1] 분류 모델 교차검증 평균 성능 결과

네 모델 중 교차검증에서 F1-score와 PR-AUC 값이 가장 높은 XGBoostClassifier를 최종 모델로 선정하였고, 테스트 세트의 성능은 F1-score 0.6988, PR-AUC 0.7685로 나타났다. 최종적으로 선정된 32개의 X 변수 데이터로 모델 구축과 미가입자 예측을 수행하였다.

3. 고객별 SHAP 분석

학습한 XGBoostClassifier 모델로 2025년 4월 기준 새로운 전체 미가입자 고객 43,892명의 월정액 서비스 가입 여부('PPM_YN')를 예측하였고, predict_proba 0.5 이상인 2,138명을 가입(Y=1)으로 예측하였다.



[그림 1] 고객 SHAP 분석 결과 예시

실제로는 미가입 상태이나 모델이 가입자로 예측한 이 고객군을 가입 추천 대상으로 선정하고 2,138명 고객별 SHAP 분석을 하였다. SHAP 분석은 모델의 예측 결과에 각 변수가 미친 영향력을 수치화하여 나타낸 것으로, 값이 클수록 가입자로 예측하는 데 긍정적인 영향을 준 것으로 해석할 수 있다[4]. 그림 1은 추천 대상 고객 한 명의 SHAP 결과를 시각화한 예시이다.

4. 고객별 맞춤형 Sales Talk 생성

SHAP 분석 결과는 서비스 가입 추천의 근거로 활용되며, OpenAI API에 전달된 입력 변수는 전체 32개 중 성별, 연령대 등 추천 근거로 사용하기 어려운 15개를 제외한 17개 변수에서 추출하였다. 고객별로 해당 17개 변수에서 SHAP 값이 양수인 변수 중 상위 3개를 선택하였으며, 선정된 변수명과 값들을 API에 전달하고 프롬프트 엔지니어링을 수행하였다[5].

모델에게 “ VOD 월정액 서비스를 권유하는 세일즈 매니저” 라는 역할을 부여하고, 전달된 변수와 값이 어떤 고객 특성을 의미하는지 이해를 돕기 위해 “ VOD 시청 시간 등급: 1~3은 시청 시간이 많은 고객” ,” 실시간 채널 시청 시간: 179 이상은 TV 시청 시간이 많은 고객” 과 같은 해석 기준 정보를 제공하였다.

또한 다양한 고객 특성 조합과 그에 따른 영업 멘트 예시를 함께 제공하는 few-shot prompting 기법을 적용하였다. 총 5개의 예시를 제시한 뒤, 추천 대상 고객의 SHAP 결과를 입력으로 사용하여 고객 특성에 맞는 맞춤형 문구를 생성할 수 있도록 하였다. 아래는 실제 생성된 Sales Talk 예시이다.

예시) SHAP 결과 키즈 가구 여부: Y, VOD 시청 시간 등급: 1, 한달 실시간 채널 시청 시간: 243시간인 고객
Sales Talk: 고객님 안녕하세요. 고객님께서 VOD 시청 시간이 많으신 점을 고려할 때, 월정액 서비스는 더 많은 콘텐츠를 경제적이고 편리하게 즐기실 수 있는 좋은 방법입니다. 특히, 키즈 콘텐츠를 무제한으로 이용할 수 있어 아이들에게도 큰 즐거움을 줄 수 있습니다. TV를 자주 시청하시는 고객님께 다양한 콘텐츠를 부담 없이 제공하는 VOD 월정액 서비스, 한번 고려해보시면 어떨까요?

Ⅲ. 결론

본 연구는 통신사의 VOD 월정액 서비스 가입률 향상을 위해 머신러닝 기반의 가입 추천 고객 예측 모델을 개발하고, SHAP 분석을 통해 예측 결과에 대한 설명력을 확보하였으며, 이를 바탕으로 대규모 언어 모델(LLM)과 연계하여 고객별 맞춤형 Sales Talk를 자동 생성하는 프레임워크를 제안하였다. 향후에는 실제 마케팅 적용 후의 고객의 전환율 분석 및 모델의 마케팅 성과 평가까지 확장될 수 있을 것이다.

참 고 문 헌

[1] 김민수 외, "설명 가능한 정기예금 가입 여부 예측을 위한 앙상블 학습 기반 분류 모델들의 비교 분석", 『한국지능정보시스템학회논문지』, 28(4), 183-200, 2022.

[2] Zhang, Y. et al., "Enhancing the Interpretability of SHAP Values Using Large Language Models", arXiv preprint, arXiv:2306.11651, 2023.

[3] Batuhan Bakirarar, "Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research", International Journal of Computer Applications, 182(1), 1-6, 2023.

[4] Scott M. Lundberg, "A Unified Approach to Interpreting Model Predictions", Advances in Neural Information Processing Systems, 30, 4765-4774, 2017.

[5] Jules White, "A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT", arXiv preprint, arXiv:2302.11382, 2023.