1. **본 론**

1) **활용데이터 정의**

본 프로젝트에서 활용된 데이터는 ‘롯데맴버스 빅데이터 경진대회’에서 주어진 롯데그룹 온라인/오프라인 계열사의 이용 이력데이터를 사용하였다. 테이블은 고객 데모 정보, 상품 구매 정보, 제휴사 이용 정보, 상품 분류 정보, 점포 정보와 엘페이 이용내역으로 총 6개의 테이블로 구성되어 있다.

데이터 활용에 있어 주어진 데이터를 모두 사용하기 보다는 선정한 주제에 부합하도록 선택적으로 활용, 고객 데모 정보, 상품 구매 정보, 제휴사 이용정보, 상품 분류정보, 엘페이 이용내역으로 5개의 테이블을 병합하여 사용하였다.

2)**파생변수 생성**

RFM(Recency, Frequency, Monetary)[1] 등급 변수 산출, RFM은 고객의 행동을 분석하기 위해 널리 사용되는 마케팅 방법론으로, 고객이 얼마나 최근(Recency)에 얼마나 자주(Frequency) 구매 했는가, 그 구매의 규모(Monetary)는 얼마인가를 기준으로 고객의 가치를 분석하는 기법이다. 세가지 요소를 통해 고객을 등급화 하게 된다. 기존의 많은 연구에서는 고객을 8 혹은 9개의 등급으로 나누는 것이 일반적이나 현재 사용되는 데이터(약 46만개)를 다분한 세트먼트로 분류할 시 고객의 행동패턴을 대변하지 못한다고 판단하여 5개의 등급으로 고객을 분류하는 RFM 변수를 도출하였다.

각 요소 점수 부여 방식은 분위수를 사용하였다. 데이터 이상치에 영향을 받지 않고 각 구간에 일정한 비율로 데이터가 포함될 수 있게 해주는 방도이기 때문이다. 각 요소를 구한 뒤, 최종 등급 산출에 있어서는. 매출기여도(Monetary)의 표준편차를 최대화 하는 가중치를 산출하고 이 가중치와 RFM 점수를 이용하여 고객별로 등급을 부여했다.

상품 중분류 정규화 변수 산출

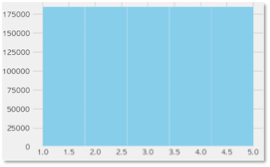
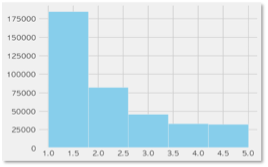
구매 상품에 따른 구매금액 비율을 기준으로 최대최소 정규화를 통해 파생변수를 생성하였다.

요일 별 인당 평균결제금액 산출

타 요일에 비해 월요일에 결제 금액이 많다는 점을 토대로 특정 요일 고객 유입이 구매로 이루어지는 고객 행동패턴이 다분할 것, 가정하에 변수를 생성하였다.

**3) 불균형 데이터 처리**

고객 구매 데이터 특성상 자사 플랫폼을 빈번히 방문해 구매하는 충성고객의 분포가 신규고객 및 잠재고객에 비해 월등히 많은 데이터 불균형이 발생, 이를 해결하기 위해 오버 샘플링 기법인 SMOTE(Synthetic Minority Over-sampling Technique)[2] 를 활용하였다. 데이터의 개수가 적은 클래스의 표본을 가져온 뒤 임의의 값을 추가하여 새로운 샘플을 만들어 데이터에 추가하는 방법이다. SMOTE를 통해 불균형 데이터를 균일한 분포의 데이터로 변화하여, 처리 전 발생할 수 있는 예측에 대한 과적합 문제를 개선하여 데이터를 활용하였다.



[기존 데이터분포] [처리 후 데이터분포]

**4) 머신러닝 모델**

랜덤포레스트(Random Forest) [3] 알고리즘은 결정트리(Decision Tree)분류기 여러 개를 훈련시켜서 그 결과들을 가지고 예측을 수행한다. 각각의 결정트리를 훈련시킬 때 배깅(Bagging), 즉, 전체 훈련 세트에서 중복을 허용하여 샘플링한 데이터셋을 개별 결정트리 분류기의 훈련 세트로 하여 훈련시키는 방식의 모델이다. 앙상블 알고리즘 중 상대적으로 빠른 수행 속도와 높은 예측 성능을 때문에 위 모델을 사용하였다.

XGBoost(Extreme Gradient Boosting)[4] 오답에 가중치를 더하며 학습을 진행하는 GBM(Gradient Boosting Machine) 모델 중 하나로 학습된 다수의 결정트리로 부터 순차적으로 트리를 학습하는 앙상블 기법인 부스팅(Boosting을 차용한 모델이다. 기존 GBM 대비 빠른 수행시간과 병렬 처리로 학습하며 분류 속도가 빠르고 과적합 규제에 기능이 있다는 점에서 위 모델을 활용하였다.

LightGBM (Light Gradient Boosting Machine)[5]은 트리 기반의 학습 알고리즘인 gradient boosting 방식의 알고리즘이다.GOSS(Gradient-based One-Side Sampling**)**기술을 활용한다. GOSS는 Information gain을 계산할 때 기울기가 작은(가중치가 작은)개체에 승수 상수를 적용하여 데이터를 증폭시킨다 데이터 분포를 많이 변경하지 않고도 훈련이 덜 된 개체에 초점을 보다 잘 맞출 수 있다. 또한 트리가 깊어지면서 소요되는 시간과 메모리 모두를 절약 할 수 있는 leaf-wise 방식을 사용한다. 이를 토대로 GBM 알고리즘들보다 압도적으로 빠른 수행 속도를 갖고 있기 때문에 위 모델을 사용하였다. .

Catboost(Categorical Boosting)[6]은 기존 GBM기반 알고리즘을 사용함에 따라 야기되는 target leakage문제, 즉 범주형 값을 수치로 바꾸는 과정에서 target value가 포함이 되어 버리는 문제를 해결하고  일괄적으로 모든 훈련 데이터를 대상으로 잔차 계산을 했던 기존의 부스팅 모델들과 다르게 일부만 가지고 잔차를 계산을 한 뒤, 이 결과로 모델을 다시 만드는 방법인 Order Boosting 방법을 사용한 모델이다. 범주형 변수에 대해 특정 인코딩 방식으로 인하여 모델의 정확도와 속도가 높기 때문에 위 모델을 사용하였다.

**5) 모델 구현 결과**

성능 평과 결과는 아래 표와 같다.

<표1>

|  |  |  |  |
| --- | --- | --- | --- |
| 모델명 | Accuracy | Logloss | Auc-Roc |
| RF | 0.994 | 0.042 | 0.999 |
| **LightGBM** | **0.932** | **0.349** | **0.973** |
| XGBoost | 0.809 | 0.647 | 0.980 |
| CatBoost | 0.847 | 0.583 | 0.966 |

\*RF – Random Forest

네 모델을 비교해본 결과 LightGBM의 결과가 세가지 평가지표에서 모두 유의미한 결과를 도출한 것을 보였다. Random Forest 모델의 경우 성능지표에서 과도하게 높은 점수가 산출된 것으로 보아 과대적합의 문제가 야기되었다고 판단해 최종모델을 LightGBM으로 선정하였다.

**6) 하이퍼파라미터 조정 후 모델 결과**

하이퍼파라미터 조정 방법으로는 랜덤서치(Random Search)를 사용하였다. 랜덤서치는 정해진 범위 안에서 난수를 생성해 최적의 하이퍼파라미터 값을 찾는 알고리즘이다. 정해진 간격(Grid) 사이에 위치한 값들에 대해서도 확률적으로 탐색이 가능하므로, 최적의 하이퍼파라미터 값을 더 신속하게 찾을 수 있다.

LightGBM 성능 평과 결과는 아래 표와 같다.

<표2>

|  |  |  |  |
| --- | --- | --- | --- |
|  | Accuracy | Logloss | Auc-Roc |
| LightGBM | 0.926 | 0.330 | 0.991 |

<표3> Classification matric

|  |  |  |  |
| --- | --- | --- | --- |
| Target | Precision | Recall | F1-Score |
| 1등급 | 0.94 | 0.98 | 0.96 |
| 2등급 | 0.93 | 0.91 | 0.92 |
| 3등급 | 0.92 | 0.91 | 0.92 |
| 4등급 | 0.92 | 0.91 | 0.91 |
| 5등급 | 0.91 | 0.92 | 0.92 |

하이퍼파라미터 값 조정 후 Logloss와 높은 예측 정확도를 보이고 있는 것을 확인할 수 있었다. 위와 같은 RFM 등급 예측치를 통해 엘페이 미사용 유저 제고를 위한 개인화 마케팅 전략 구상에 있어 신뢰성을 갖고 모델이 활용 될 수 있다고 판단 된다.

참고 문헌

[1] https://zephyrus1111.tistory.com/12

[2] smote-oversampling-for-imbalanced classification/

[3] <http://www.incodom.kr/Random_Forest>

[4] <https://seethefuture.tistory.com/91>

[5] <https://kicarussays.tistory.com/38>

[6] https://dailyheumsi.tistory.com/136