

Study for the Presenting personalized marketing strategies for Lotte members' customers through LPAY rating prediction with Machine Learning and Data Analysis

CUAI 5 기 데이터 분석 7 팀

강영훈(경영학), 김소은(통계학), 원민재(경영학),

[요약]Machine Learning 과 데이터 분석을 통해 Lpay 미사용 유저의 Lpay RFM 등급 예측을 통해 잠재 고인 유인을 위한 개인화 마케팅 전략 도출을 진행하고자 함. LightGbm 모델을 통해서 표본 중에서 2.29% 예측을하여 실제 업무 적용시 480,900 명을 대상으로 Lpay 전환을 위한 마케팅 전략을 제시함.

1.서 론

코로나 19 및 스마트 폰을 통한 비대면 금융거래의 보편화 현상은 현재 간편 결제 시장의 성장을 불러일으켰다. 실제로 한국은행의 보도자료에 따르면 체크카드를 포함한 간편결제 서비스의 이용현황에서 이용 건수는 2016 년 210 만건 대비 2020 년 1454 만건으로 1244 만건 증가하였고, 이용금액 또한 645 억원에서 4492 억원으로 3807 억원 증가하였다.



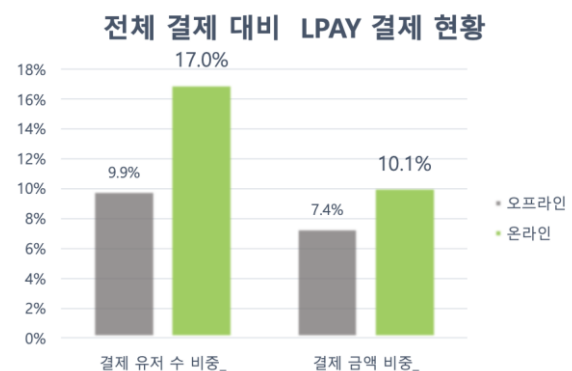
[그림 1] 간편결제 서비스 이용 현황

해당 간편결제 시장의 성장에 따라서 기존의 유통 산업에서도 '특정 생태계를 조성해 재화나 서비스

이용 시 다른 선택을 제한하게 종속 시켜 재구매를 촉진시키는 락인 효과 및 빅데이터 확보를 통한

신사업 확장을 위한 간편결제 서비스를 확대하고 경쟁력을 갖추려는 시도를 하고있다.

롯데멤버스 또한 L.Pay 간편결제 서비스 런칭 및 금융서비스(썸뱅크)와의 연계를 통해 충성 고객 확보 및 서비스 경쟁력 확충 등의 사업 확장을 진행하고 있으나, 해당 서비스의 경쟁력이 동종 업계 대비 저조한 편이며 (네이버, 이베이 코리아, 11 번가, 쿠팡, 신세계, 롯데 6 개 기업중 6 위) 유통/제휴사 결제 이력이 있는 유저 중 LPAY 결제 이력이 없는 유저가 대다수이고 온라인 매체에서의 LPAY 결제 금액 비중이 현저히 낮은 편이다.



[그림 2] 전체 결제 대비 LPAY 결제 현황

현재 Lpay 결제 이력 유무에 따른 고객별 월 온라인 ARPU 는 50,000 원 이상 차이가 발생하여, Lpay 결제 이력 유무에 따라 온라인 매출액 증대가 예상되며 실제로 Lpay RFM 등급이 높은 유저일 수록 결제 제휴사 / 유통사 수가 많은 것을 확인할 수 있다. 따라서 우리는 이러한 LPAY 결제의 활성화를 통해 ARPU 증대, 락인효과 등을

촉진시키기 위해 머신러닝과 데이터 분석을 통해 개인의 LPAYRFM 등급을 예측하고 이에 대응한 개인화 마케팅 전략을 제안하고자 한다.

2. 본 론

1) 활용데이터 정의

본 프로젝트에서 활용된 데이터는 '롯데멤버스 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 6 개의 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 5 개의 테이블을 병합하여 사용하였다.

2)파생변수 생성

RFM(Recency, Frequency, Monetary)[2] 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 9 개의 등급으로 나누는 것이 일반적이거나 현재 사용되는 데이터(약 46 만개)를 다분한 세트먼트로 분류할 시 고객의 행동패턴을 대변하지 못한다고 판단하여 5 개의 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.

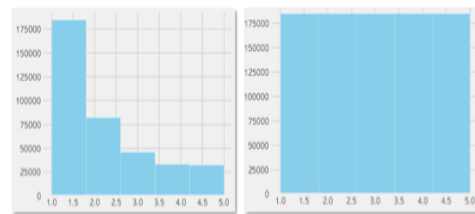
상품 중분류 정규화 변수 산출
구매 상품에 따른 구매금액 비율을 기준으로 최대최소 정규화를 통해 파생변수를 생성하였다.

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)[3] 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.



[그림 3] 기존 데이터분포(좌), 처리 후 데이터분포(우)

4) 머신러닝 모델

랜덤포레스트(Random Forest) [4]
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)[5] 오답에 가중치를 더하며 학습을 진행하는 GBM(Gradient Boosting Machine) 모델 중 하나로 학습된 다수의 결정트리로 부터 순차적으로 트리를 학습하는 앙상블 기법인 부스팅(Boosting)을 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)[6]은 트리 기반의 학습 알고리즘인 gradient boosting

방식의 알고리즘이다. GOSS(Gradient-based One-Side Sampling)기술을 활용한다. GOSS 는 Information gain 을 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)[7]은 기존 GBM 기반 알고리즘을 사용함에 따라 야기되는 target leakage 문제, 즉 범주형 값을 수치로 바꾸는 과정에서 target value 가 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.

3.결 론

1) 모델 구현 결과

성능 평가 결과는 아래 표와 같다.

〈표 1〉

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 LightGBM 의 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 LightGBM 으로 선정하였다.

2) 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 평가 결과는 아래 표와 같다.

〈표 2〉

	Accuracy	Logloss	Auc-Roc
LightGBM	0.926	0.330	0.991

〈표 3〉 Classification matrix

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

하이퍼파라미터 값 조정 후 Logloss 와 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.

3) 개인화 마케팅 전략 제안

롯데멤버스사의 Lpay 결제 현황은 전체 결과의 80%가 전체 원인의 20%에서 일어나는 현상인 파레토 법칙 (80-20 rule) 에 따라서 기존 Lpay 사용자 중 Lpay RFM 1 등급 유저의 매출액이 총 매출이 75 % 를 차지하고 있다. 따라서 LPAY 미사용 유저 중 1 등급으로 예측되는 유저를 대상으로 Lpay 전환 마케팅이 우선시 되어야 한다고 보았다.

〈표 4〉 Lpay 등급별 매출액

L_등급	L_PAY 총 매출액	매출액 비중
1 등급	15,993,060,616	75%
2 등급	3,288,716,265	15%
3 등급	1,338,420,026	6%
4 등급	578,708,224	3%
5 등급	227,855,300	1%

RFM 1 등급 예측 유저의 전환을 위해 기존 Lpay 1 등급 유저의 소비 회수가 높은 상품의 중분류 카테고리 상품을 미끼 상품 (Loss Leader) 로 활용해 Lpay 결제 전환을 제고하는 마케팅 전략을 제시하였다. 기존 1 등급 유저의 선호 상품 중분류는 채소, 대용식, 과자, 유제품, 과일 순이었고 이를 첫 Lpay 로 결제시 990 원 특가로 제공, 해당 이벤트를 1 등급 예측 유저에게는 푸시 메시지로 발송, 최초 결제시 할인을 제공하는 쿠폰을 발급하는 등의 마케팅 전략을 제시하였다.



[그림 4] 마케팅 전략 사례

또한 Lpay RFM 의 등급별 특징을 통해 유저 등급을 부여한 이후 차등적인 프리미엄 서비스를 제공하는 것을 제안하였다. Lpay 등급 별로 다음과 같은 특징을 가지고 있는데, 이러한 특징을 프리미엄 멤버십의 등급제에 반영하여 차등적인 쿠폰을 지급하고 해당 특성을 바탕으로 해당 지표들을 Lpay 결제 상승의 Aha Moment 로 활용할 수 있다.

<표 5> Lpay 등급별 결제 추이

Lpay_등급	월별 평균 결제 일자	월별 평균 결제금액	월별 평균 결제 횟수
1 등급	3.54	150818.11	9.35
2 등급	2.62	117667.44	7.70
3 등급	2.22	100470.31	6.17
4 등급	2.07	92084.96	5.76
5 등급	1.91	87788.83	5.70

5) 기대효과

롯데멤버스의 제공 데이터의 Unique 고객 수는 총 29,874 명이고 LPAY 미가입 유저 중 LPAY 1 등급 예측 유저 총 683 명이다. 따라서, 제공 데이터 기준으로 1 등급 예측 유저 / 전체 표본=

$683 / 29,874 = 2.29 \%$ 이다. 이와 같은 수치를 롯데 멤버스 홈페이지에서 발췌한 롯데멤버스의 연간 이용회원 수에 적용하면 롯데 멤버스 전체 이용 유저 = 2,100 만 명이고 전체 유저 X 1 등급 예측 유저 비율 = 2,100 만명 X 2.29% = 480,900 명이다. 따라서 1 등급 예측 유저 LPAY 전환 시 기존 LPAY 가입 유저 대비 9.61 % 확보 효과가 발생할 것으로 추산한다. 또한 롯데 멤버스 제공 데이터 기준 고객 당 월 평균 소비금액을 Lpay 전환 시 기대되는 Lpay 월 전환 금액으로 추산하면 전체 1 등급 예측 유저 수 X 전환 시 기대되는 LPAY 월 전환 금액 = 480,900 명 X 298,847 (원) 으로 총 143 억 가량이 Lpay 전환 결제액으로 증가할 것으로 추산할 수 있다.

또한 기존 0 등급 (Lpay 미 이용) 유저의 제휴/유통사 이용수가 2~4 회이고, 기존 1 등급 유저의 제휴/유통사 이용수 가 6~8 회인 점을 고려하였을 때, LPAY 전환 시 평균적으로 2~4 개의 제휴사 / 유통사를 추가로 이용하게 유도하는 효과를 발생하여 충성 고객 확보에 용이할 것이다.

결론적으로 LPAY 미가입 유저에 대한 예측 등급별 타겟팅 마케팅을 이용한 가입 전환을 통해서 LPAY 결제액과 매출액 및 충성 고객 수 상승에 기여 할 수 있다.

참고 문헌

- [1] 간편결제 서비스 이용 현황.Avaliable: <https://www.yna.co.kr/view/GYH20210329001900044>
- [2] RFM <https://zephyrus1111.tistory.com/12>
- [3] SMOTE <https://machinelearningmastery.com/smote>
- [4] Random Forest http://www.incodom.kr/Random_Forest
- [5] XGBoost <https://seethefuture.tistory.com/91>
- [6] LightGBM <https://kicarussays.tistory.com/38>

[7] Catboost
<https://dailyheumsi.tistory.com/136>