목차

1. RFM모델

2. 데이터 사전 분석(이상 유저 등)

3. 모델 기준 수립

5. 전처리 시행

6. 그룹별 분포도

7. 결론

8. 그 외

1. RFM모델

Recency는 고객이 얼마나 최근에 구매했는지,

Frequency는 고객이 얼마나 자주 방문했는지,

Monetary는 고객이 돈을 얼마나 소비했는지를 의미합니다.

각 피쳐별로 우수고객과 잠재고객을 분류해 고객군을 나눠 그룹별 마케팅을 시행하는 방식을 말합니다.

피쳐별로 구획(점수)를 나누어 등급을 매기며 만약 피쳐별로 4개씩 나눈다 가정하면 64가지의 고객 군이 나오게 됩니다. 이들 군 중 조건에 부합하는(마케팅 목적에 맞는) 유저군을 선택해 맞춤형 마케팅을 시행 할 수 있습니다. 다만 단점으로는 구간이 하나 하나 늘때마다 고려해야 할 그룹수가 지나치게 많아진다는 단점이 있습니다.

그래서 다른 방법으로는, 개략적인 정보를 얻고자 한다면 각 구간별로 점수를 매겨 총합점수를 새로운 변수로 삼아 우수고객과 잠재고객을 분류하는 방법도 있습니다. 다만 단점으로는 피쳐별 세분화가 정확히 반영되지 않아 세부적인 마케팅 방향의 근거로 삼기에는 부정확 할 수 있습니다.

따라서 두가지 방법 중 무엇을 쓸지, 그리고 구획을 몇 개씩 나눌지는 데이터를 확인해 본 이후에 판단하려 합니다.

1. 데이터 사전 분석(이상 유저 등)

피쳐들의 분포도 등을 개략적으로 살펴보고 난 뒤, 모델의 세부사항들을 결정하려 합니다.

또한, Recency의 기준을 결정하기에 앞서, 과제표에 의하면 21년 4월 19일까지의 결제 데이터를 보유한 셈이므로 기준일을 4월 20일 00시로 가정하고자 합니다. 왜냐면 현재 날짜를 기준으로 집계할 경우 1달 반동안 모인 거의 모든 데이터가 최신 결제기록이 1년전으로 표현되어 왜곡될수 있기 때문에, 분석 날짜를 데이터가 모인 직후인 4월 20일로 가정하였습니다.

유저 데이터는 테이블 하나에 전부 들어있으며, 플레이 해본 경험과 상품 개요 데이터 차트에 대해 간략히 보면, 인게임 재화 코인, 퍼즐 게임 클리어에 도움이 되는 소모형 아이템 해머 다트 등이 주력 상품이며 12시간등 기간제 아이템도 구매할 수 있습니다. 재화의 가치는 조금씩 다르겠지만 일반적으로 번들에는 같은 수의 소모형 아이템이 들어있습니다. 다만 로켓과 TNT는 별도 번들이나 기간제로 구할 수 있습니다. 혹은 코인을 소모해 섞어넣는 식으로 코인의 지속적인 소모를 유도하고 있습니다.

인게임 튜토리얼 당시 로켓으로 화려하게 퍼즐을 날려버리는 모습을 직접 겪게 해줌으로써 아이템이 퍼즐 클리어에 큰 도움이 되는 상품임을 경험시키고 구매를 유도하게끔 하는 것은 좋다고 생각합니다.



유저 데이터로 넘어가서, 데이터 셋의 크기는 3379행. Event\_timestamp는 UTC시간이므로 전환해서 사용해야 할것같습니다. Custom\_user\_id는 유저 고유 ID를 의미하는데 유니크값이 531개이므로 해당 기간동안 결제한 총 paying user는 531명. 결제기록만 존재하기에 UV는 알 수 없습니다. 자연히 구매자 전환 비율(PUR)이나 이용자당 평균 매출(ARPU) 역시 확인 할 수 없습니다.

유저 ID가 1번부터 부여되었다면 알기 쉬웠겠지만, ID 최소값이 4830인걸로 보아 섣불리 추측할 수 없어 포기하였습니다.

* 이상치 탐색

사전에 해킹유저나 분석결과를 왜곡할 수 있는 이상치의 존재를 언급하셨기에 확인해 보았습니다. 실제로 요약표에서도 확인 할 수 있듯이 가격이 누락된 데이터가 5건 확인되었고 이는 삭제처리 했습니다.

판매 위치와 상품이 매치되지 않는 경우가 있는지 확인해 본 결과, 상품 정보에서 팝업으로 기록된 상품들 역시 상점에서 구매하므로 유저 데이터에는 shop으로 기재되어 있습니다. 즉, 유일한 예외로는 배너 구입으로만 구할수 있는 vip\_pass 뿐입니다. 그런데 상점에서 구매한 vip\_pass도 존재합니다.

* 최중요 보고 사항(시간의 차이)

해당 유저 결제 정보의 시간의 축이 매우 뒤틀어져 있습니다. 과제 개요표에는 [데이터는 2021년 3월 9일부터 2021년 4월 19일까지 인앱 결제 데이터의 일부와 상품 정보 데이터 입니다.] 라고 기록되어 있으며 강조 표시까지 되어 있고, 해당 테이터의 event\_date 칼럼 역시 이에 맞게 기록되어 있습니다. 그런데 event\_timestamp칼럼을 변환해보면 UTC로는 시간 이하는 맞는듯 하지만 달이 무려 21개월씩의 차이를 보입니다.



이에 확인해본 결과, event\_date의 분포 자체가 이상함을 확인했습니다. 해당 칼럼의 경우 2021년 3월 9일부터 3월 31일까지, 그리고 **2023**년 4월 1일부터 19일까지 기록이 되어 있었습니다.

event\_date의 신뢰도는 사실상 없는것과 마찬가지임을 감안하여 해당 칼럼은 **완전히 무시하고**, 적어도 분포 자체는 고르게 되어있는 event\_timestamp를 임시로 기준으로 삼되, 문제에서 제시했던 기한인 20210309-20210419에 부합하도록 21개월의 시차를 적용해 적용하였습니다.

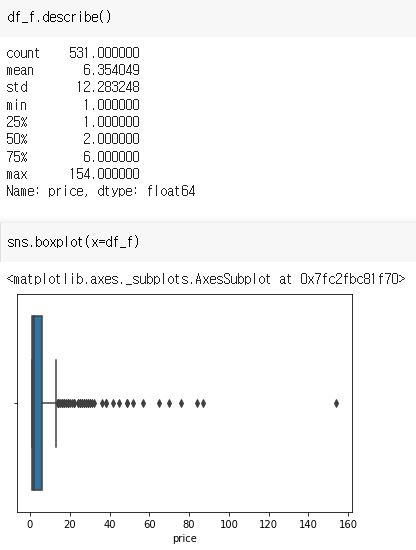
결제 화폐가 달러이고 시간대의 분포를 감안하여 KST대신 UTC로 유지하였습니다.

추가)

제출 직전 발견한 결과 결제이력 파일의 이름이 20221209\_20230119\_ZJ\_IAP\_data 임을 미뤄볼 때, 처음부터 파일에 잘못된 두개의 시간이 섞여있던 오류가 있는게 아닐까 생각됩니다.

1. 모델 기준 수립

유저 계정인 custom\_user\_id를 기준으로 그룹핑하여 상세정보를 조회해본 결과, 세 그룹으로 나누기로 결정했습니다. RFM의 각 기준에 따라 분류했을 때 분포도가 비슷했다는 점과, 상위 25, 50, 75%의 수치들을 미뤄 볼 때 0~25%, 25~75%, 75%~ 의 세 그룹으로 나누는게 가장 성격을 잘 반영한다고 보았기 때문입니다.

평균을 이용하지 않은 것은 분포도 때문이었는데, 가령 결제 빈도수를 본다면 전체 531명의 유저분들 중 평균 결제수는 6회에 달했지만, 가장 많이 결제하신 손님이 154회나 결제하는 등 평균값 자체의 신뢰도가 높지 않았기 때문입니다.

옆 자료와 같이, 허용 범위 내에서 최대값은 10, 상위 75%에 해당하는 유저분은 6회의 결제 빈도수를 기록하고 있었습니다. 따라서 75%정도면 일반적인 거래횟수라고 보아도 무방하다고 판단하였습니다.

마찬가지 이유로, 하위 25% 이하는 단 한번 결제한, 수치로도 비율로도 다회 결제와는 관련이 없는 수치라고 보았기 때문입니다.

또한, 소수점의 경우 반올림값을 사용하였습니다.

이에 따라 나누게 되는 스코어는 다음과 같습니다.

|  |  |  |  |
| --- | --- | --- | --- |
| Score | Recency | Frequency | Monetary |
| 고득점 특징 | 최근에 결제할수록 | 자주 결제할수록 | 총 금액이 많을수록 |
| 3 | 5일 이하 | 6회 초과 | 24달러 이상 |
| 2 | 5일 초과 27일 이하 | 1회 초과 6회이하 | 3달러 초과 24달러이하 |
| 1 | 27일 초과 | 1회 이하 | 3달러 이하 |

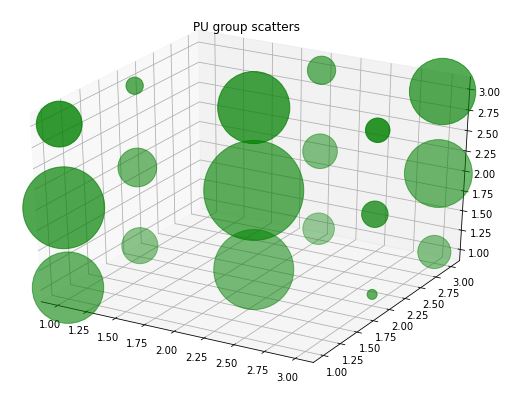
1. 전처리 시행

설정한 기준에 따라 데이터프레임을 구해 각 유저별로 점수를 매겨 점수표를 만듭니다.



1. 그룹별 분포도





1. 결론

분석을 시작하기에 앞서, 전체 531명의 플레이어중 최소 3점, 최대 9점을 기록하는 분포도상, 또 대부분의 20-80법칙등을 고려하면 낮은 점수대가 많은 분포를 기록할것이라 생각했습니다. 하지만 생각보다 3,3,3으로 최근결제함, 일정 금액 이상 구매함, 자주(6회 넘게) 구매해주신 VIP고객님이 44명으로 약 10%나 차지하는 것은 꽤 놀랐습니다. 이는 한번 구매한 고객님이 지속적으로 구매해주고 계시다는 의미일수 있습니다.

더욱이 최우수 고객님이라고 볼수 있는 상위 득점자 2,3,3 3,2,3 3,3,2 또한 46명, 8명, 6명을 기록하는데 이것이 시사하는 바는 꽤 큽니다. VIP까지 합한다면, 약 100여명으로 20%가

즉 결제를 해주고 계시는 열성적인 분들의 경우 충성도가 높다는 의미이기 때문입니다.

다만 233이 다른 조합에 비해 수가 압도적으로 많다는 점 또한 주목할만 합니다. Recency\_score 는 5일일 기점으로 나뉘는데, 소모아이템이 많고 플레이에 열심힌 유저들분조차 주에 1회 이하로 구매를 하고있다는 뜻일수 있기 때문입니다.

평균 점수대는 많을 수밖에 없었습니다. 중간그룹인 일반 그룹이 각 피쳐별로 50%씩 영역을 차지하고 있었기 때문입니다.

111을 기록한 10%대의 손님들은 구매한지 오래전에 게임을 떠나신 고객층이라 생각됩니다. 또한 Recency\_score가 1점대인 즉 1번만 구입하신 고객님들은 고가의 패키지를 구매하지 않는 경향이 뚜렷하며 Monetary\_score = 1을 기록하신 세트로 그룹화 된 고객님들이 많이 보입니다. 실제 집계에서도 1.99달러 짜리인 warm\_welcome\_bundle을 구매하시는 분들이 357건으로 전체 3374건중 10% 이상을 차지했습니다.

마찬가지로 1.99달러인 dont\_giveup\_now1 팩도 313건으로, 둘중 하나는 반드시 사고 시작한다고 봐도 무방하다고 생각합니다.

1. 이슈, 특이사항, 기타

처음 전처리 단계에서 날짜 이슈가 문제가 되어 굉장히 당황했었지만, 이것도 시험의 일환이라고 가정하고 임의로 해석, 풀이 하되 그 근거를 제시하는 방식으로 진행하였습니다.

과제와는 관계없는 게임을 몇시간 해보며 느낀점이지만, 상점 탭에 아이템을 오래 클릭하면 설명이나 사용 영상이 나온다던가 하면 좋겠다는 느낌도 받았습니다.

이 외에 이번에는 상품별 소모기간, 특히 시간제 아이템이 어느 타이밍에 많이 쓰이고 활성화 되는동안 얼마나 유효한 플레이가 이루어지는지, 주기적으로 구입되는 상품이 있다면 어떤것들이 있을지도 들여다 보고 싶습니다.

고객지표 수식을 세워 등급별로 세분화 시켜보고 RFM 피쳐별 계수값이 실제 현장에선 어떤 식인지, 그리고 어떤 과정을 통해구해지는지도 알고 싶습니다.