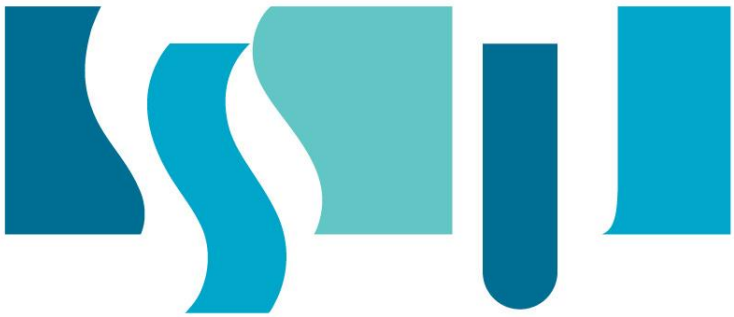
2023 년도 1 학기 전공 종합 설계 논문

RFM을 통한 고객 세분화 및 이탈 고객 예측



숭실대학교

산업정보시스템공학과

20182482 박성복

20190552 손지영

20202850 김규리

20201341 원지수

목 차

제 1장 서론

1.1 연구 배경 및 목적

CRM(Customer Relationship Management: 고객 관계 관리)이란 고객에 대한 정보를 수집하고, 수집한 정보를 분석한 후 효과적으로 활용함으로써 고객을 적극적으로 관리하고 유지하며 고객의 가치를 극대화시키기 위한 기업 마케팅 전략이 IT 기술과 결합한 것이다. 이를 통해 충성 고객의 유지 비율을 향상시킬 수 있고, 고객의 이탈로 인한 손실을 최소화할 수 있다. 또한, 주요 고객의 로열티 강화는 고객 당 거래건수, 거래단가, 거래기간 등을 증가시키며, 핵심 고객과 유사한 속성을 지닌 잠재 고객을 선별하여 신규 고객으로 유인함으로써 수익을 극대화할 수 있다. 성공적인 CRM은 고객 관리뿐 아니라 기업가치 창출이 증대하며, 협력업체들 과의 원활하고 지속적인 관계를 유지하는데 도움을 준다.

경제와 IT의 발달로 인해 고객의 구매력은 커지고 다양한 상품을 손쉽게 접할 수 있어 기업이 고객의 개성에 맞는 서비스나 제품을 제공하지 못한다면, 기존 고객들의 이탈을 막기 어렵고, 휴먼 고객이 될 가능성이 크다. 이탈이 일어나기 전 고객은 서비스 활용 저하, 제품에 대한 불만 표현, CLS(Customer Life Stage: 고객의 생애 단계)의 변화와 같은 이탈 징후를 보여준다. 고객이 이탈하기 이전에 미리 예측하고 대비하는 것의 필요성이 증가함에 따라 이탈 예측 모형에 대한 연구도 증가하고 있다. Jesper은 일정한 기간을 기준으로 이탈 여부를 결정하고 이를 LSTM을 통해 고객의 이탈을 예측하였다. 그러나 현실에서는 고객 이탈의 정의가 기업, 고객 개개인 마다 다르다는 문제점을 가지고 있다. (2017, Jesper Ljungehed)

기업은 위와 같은 이탈 징후 및 데이터를 정확하게 이해하고 예측 모형을 통한 분석을 이탈 방지 활동으로 전환할 수 있어야 한다. 이탈 방지 활동 중 가장 대표적인 것은 RFM이나 CLV(Customer Lifetime Value: 고객 생애 가치)을 이용하여 고객들을 등급화 하는 것이다. 등급에 의해 분류된 고객들에게는 등급별 다른 마케팅 전략을 적용해야 하며, 이를 통해 기업은 경쟁 우위를 확보할 수 있다.

본 연구에서는 비계약 업종의 구매 데이터에 대하여 목적에 알맞게 변형된 RFM을 적용하여 고객을 5가지 그룹으로 분류한 고객 세분화 방법을 제안하였다. 그 후, 모델을 통하여 각 그룹에 대한 이탈 가능성을 분석하고 그에 따른 맞춤 마케팅을 적용하여 충성 고객의 비율을 유지 및 증가하는 방법을 제안하고자 한다.

1.2 이론적 배경 선행 연구

1.2.1 고객 세분화

1.2.1.1 고객 세분화 의의

고객 유형화 또는 고객 세분화(Customer Segmentation)는 시장 세분화의 다양한 기준 중 고객에게 초점을 맞추어, 기업의 주된 소비자 집단을 특성에 따라 군집화해 고객 중심적 접근을 가능하게 한다. 시장 세분화는 세분시장의 규모를 비롯한 세분시장의 요건을 규명하여 목표시장을 확인하는 데에 목적이 있다. (박지현, 2022) 이러한 작업은 흔히들 데이터베이스마케팅 기법을 이용하여 자사 고객의 실적에 따라 여러 등급으로 분류되며, 대고객 전략 도출 및 접근방법에 대해 고민을 하기 위한 도구로 활용된다. (2011, 오준수)

세분화 이론은 지속적인 진화의 과정을 겪어 왔다. '전통적 세분화(Traditional Segmentation)'은 상품에 대한 욕구 및 구매자 행동에 근거하여 고객들을 분류하는 것으로 정의된다. 군집분석이나 판별분석과 같은 다양한 다 변량 통계기번을 이용하여 행동 패턴이나 데이터의 특징이 유사한 고객들을 분류해왔으며, 이렇게 분류된 고객들에게 상이한 제품을 제안하거나 다이렉트 마케팅 캠페인에 활용해 왔다. (2006, 윤종욱) 하지만 최근 들어 이러한 '전통적 세분화'는 많은 비판을 받게 된다. (1993, Peppers and Rogers) 정적인 지식을 활용한 기존 세분화 기법은 시간이 경과하게 되면서 쓸모가 없어지게 되었으며, 군집내 모든 고객에게 동일한 마케팅을 적용할 수 없다는 문제점이 제기되었다. 이러한 문제점은 고객의 동적 특성을 반영한 고객 세분화 방법을 통해 보완되었고 최근 들어 많은 연구가 진행되고 있다. (2003, 이재신)

이러한 고객 세분화는 기업이 소비자의 특성을 고려하여 유형화하고, 각 군집에 적절한 마케팅 및 고객 중심적 사고가 가능하게 해주므로 산업 분야, 특히 마케팅 서비스업을 중심으로 다양한 연구가 진행되고 있다. 유통 분야에서는 백화점 고객의 카드 사용 내역으로 매출, 구매 건수, 구매일 수를 통한 행동 특성들을 파악하여 의사결정 나무 분석을 활용하여 타겟 고객 세분화하여 파악하였다. (2010, 채경희/김상철) 항공사 이용객의 구매 내역(마일리지)를 기반으로 고객 생애가치를 계상하고, 기존 분류에서 더 심화적으로 세분화하기도 하였다. (2014, 박광식) 또한, 외래 환자 고객을 대상으로 이용기간, 총 방문횟수, 평균 진료비를 파악하여 세분화 계층을 분류한 경우도 있다. (2004, 이소영 외 3)

1.2.1.2 고객 세분화 접근법

기본적인 고객 세분화 방법에는 인구 통계학적 접근법, 심리 통계학적 접근법, 행동 특성 접근법이 있다. 인구 통계학적 접근법(Demographic Approach)은 연령, 성별, 소득 등의 고객의 외형적 특성을 기반으로 고객을 유형화하는 것이다. 많은 데이터를 쉽게 얻을 수 있다는 장점이 있지만, 단순한 특성들이기 때문에 고객의 심리를 파악하는 데 어려움이 있다. 심리 통계적 접근법(Psychographic Approach)은 상품이나 브랜드에 대한 태도와 구매 등에 있어 가시적이지 않은 인간의 심리적인 부분을 파악하여 고객을 분류하는 것이다. 그 예로는 고객의 라이프스타일,

신념, 개성, 사회적 지위 등이 있다. 이러한 심리적인 요소들이 단순한 인구 통계학적 세분화를 보완해주었지만, 수집이 어렵고 주관적인 결과이므로 신뢰도가 떨어진다는 단점이 있다. 마지막으로 행동 특성 접근법(Behavioral Approach)은 소비자의 행동에 기반을 둔 분석으로 소비자의 추구편익, 사용률, 사용 경험, 구매자의 상태 등이 있다. 분석에 필요한 데이터만 축적되어 있거나 접근이 가능하면 기계 학습의 도구를 활용하여 다양한 방법으로 분석할 수 있다.

1.2.1.3 RFM 기법

RFM은 고객의 행동을 분석하기 위해 널리 사용되는 마케팅 기법으로, 고객이 얼마나 최근(Recency)에 얼마나 자주(Frequency) 구매했는가, 그 구매의 규모(Monetary)는 얼마인가를 기준으로 고객의 가치를 분석하는 마케팅 기법이다. (2017, 지현정 외 3) 최근성은 고객이 최근에 구매한 고객일수록 앞으로도 구매할 가능성이 높다는 판단을 통하여 높은 점수를 부여한다. 빈도성은 일정 기간 동안 많이 구매한 고객일수록 앞으로도 구매할 가능성이 높다는 판단 하에 높은 점수를 부여한다. 총 구매액은 큰 구매의 규모일수록 앞으로도 기업에 많은 돈을 사용할 것이라는 판단하여 구매액이 클수록 높은 점수를 부여한다. 일반적으로 전통적인 RFM 모형(1994, Hughes)을 따라 5개의 균등한 집단으로 나누어 점수를 부여하여 총 125개의 군집을 형성하고 고객을 세분화하였다. 이러한 방법은 기업에 이익을 주는 우수고객을 가려내는 것에는 효과적일 수 있으나, 그 외의 일반 고객이나 이탈고객에 대한 파악은 어렵다. 즉 군집 수가 너무 많아 실질적으로 고객 관계 관리 등 마케팅 활동에 활용하기에는 어려움이 있다. 또한 근본적으로, Hughes (1994)는 고객을 각 기준에 따라 5개의 집단으로 나누는 것에 대한 근거를 제시하지 못했다는 한계가 있다(류귀열, 문영수, 2013)

도희정(2007)은 백화점 데이터를 이용하여 기존 고객들의 특성을 통해 신규고객을 세분화하고 미래의 구매력을 예측하는 과정에서 RFM 기법에서 착안한 CFM(Continuation, Frequency, Monetary)를 통하여 점수를 부여하고 이를 기준으로 각 기간별로 K-means clustering을 통하여 고객 세분화를 수행하였다. 이는 시간의 변화에 따른 고객 세분화가 이루어질 수 있으며 회사의 충성 고객그룹 파악 및 정확한 고객들의 구매 패턴 변화를 파악한 연구라는 점에서 의의가 있다.

본 연구는 도희정(2007)의 연구에서 착안하여, 본 연구의 데이터에 알맞은 RFM을 적용하고, 기간별 클러스터링을 통한 고객 세분화를 하여 다양한 군집화 알고리즘에 적용해보고자 한다. 이를 통하여 기간별 이탈 여부 및 구매 패턴을 파악하고자 한다.

1.2.2 고객 이탈 예측

고객 이탈예측은 머신 러닝 기반의 주요 CRM 연구 주제 중 하나이다. (2012, Kim and Shin) 경쟁이 치열한 환경에서 효과적인 이탈 예측은 CRM뿐만 아니라, 현재 기업에게 돈이 되고 앞으로 돈이 될 가능성이 있는 고객을 정확하게 이해하고 이를 기반으로 마케팅, 영업, 서비스 활동을 전략적으로 추진하는가를 결정하는데 중요한 역할을 하므로 이를 성공적으로 예측하기

위한 모형개발이 많이 이루어져 왔다. 과거에는 전문가 시스템 등의 사람이 만들어 놓은 지식을 기계에게 직접 주입하는 방식으로 구현하였다. 이러한 방식은 다양한 상황에 대응 가능한 지식 개발의 어려움 등으로 제한적으로 활용되었다. 기술이 발달함에 따라 의사결정나무, 인공신경망과 같은 단일 알고리즘 학습 및 성능 평가를 하는 단계를 거쳐 최근에는 앙상블 모형이나 이종모형을 연결한 하이브리드 모형을 개발하려는 시도가 많아졌다. (2009, Oh et al) 이탈 예측의 정확도를 높이고자는 것이 머신 러닝 기반 이탈예측 연구의 공통적인 목적이며, 이는 크게 두 가지의 연구 방향으로 구분된다. 첫 번째는 하이브리드 모델을 활용하여 고성능의 이탈예측 모형을 구축하려는 것이고(Oh et al., 2018), 두 번째는 분석에 사용되는 데이터의 불균형 문제를 개선하거나, 예측성능에 영향을 미치는 아웃라이어를 효과적으로 제거하는 등 모형 개발 이외의 프로세스 적인 차원의 개선을 이루려고 하는 연구이다(2009, Tsai and Lu). 대표적인 선행 연구로는 보험 가입자들의 인구 통계학 및 정책을 이용하여 인공신경망을 통해 이탈고객을 예측(2001, Yeo), 금융권 고객들의 시계열 정보를 사용하여 의사결정나무와 인경신경망을 통한 이탈고객 예측(1998, Eiben)이 있다.

Mehdi는 앞의 RFM을 적용하여 이탈 예측을 하였는데, 각 그룹의 RFM을 통하여 고객을 4가지 유형으로 군집화 한 후, 의사결정나무 모델을 통하여 각 군집의 규칙을 발견하였다. (2017, Mohammadzadeh) 그 결과 AUC 기준 약 0.97로 이탈 고객을 분류하는데 성공하였으며, 이는 RFM이 CRM과 함께 이탈 고객 관리를 할 수 있음을 보여주었다. 또한, 고객의 1년간 데이터의 패턴 경향을 더하여 머신 러닝을 한 결과 고객의 잠재 행동 변화를 더 잘 예측하였다. (2023, Zelenkov) 이는 시계열 적인 요소가 고객의 잠재 행동 변화에 영향을 주며 이탈 여부에도 영향이 있음을 알 수 있다.

모델을 통하여 얻어진 정보를 활용하여 기업 활동의 문제점을 평가하고 적절한 개선이 이루어진다면 고객 이탈은 기업에게 잠재적인 위협이 되지 못할 것이며, 지속적인 가치 제공을 하여 고객과의 관계유지 통해 고객이 기업에 대해 지닌 잠재적인 가치까지 흡수할 수 있을 것이다.

제 2장 본론

2.1 프로세스 설명

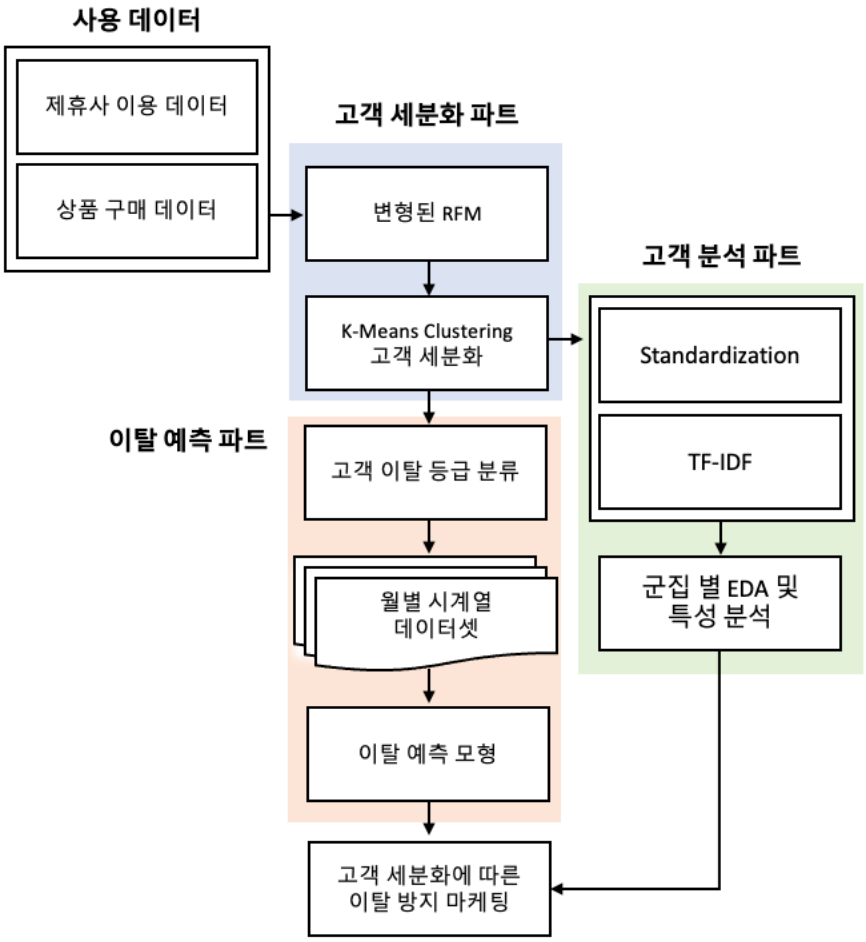
2.1.1 데이터 설명

본 연구에서 사용된 데이터는 롯데멤버스가 주관하는 제 7회 롯데멤버스 빅데이터 경진대회에서 제공받은 데이터로, 롯데그룹 계열사의 온라인, 오프라인 이용 이력이다. 고객 데모 정보, 상품 구매 정보, 제휴사 이용 정보, 상품 분류 정보, 점포 정보, 엘페이 결제 내역 데이터를 포함하고 있다. 여기서 제휴사는 엘포인트와 엘페이가 제휴되어 있는 롯데 그룹사를 의미하며, 유통사는 제휴사 중 상품을 구매할 수 있는 유통 업종 제휴사 중 일부이다. 상품 구매 정보와 제휴사 이용 정보 간에는 제휴사가 중복되지 않는다. 데이터 관찰 기간은 1년이며, 2021년 1월 1일부터 2021년 12월 31일까지 수집된 데이터가 사용되었다. 개인 정보 보호를 위하여 개인 신상 정보 및 개인을 식별할 수 있는 정보는 데이터 수집 과정에서 제거되었으며, 구체적인 거주지, 점포, 제휴사 정보는 데이터 비식별화를 위해 암호화되어 제공되었다.

2.1.2 프로세스 구성

본 연구는 고객 차별화의 일환으로, 시변성을 가지고 있는 고객의 구매 이력을 이용하여 고객을 세분화하고 등급을 부여하는 새로운 방법을 제안함으로써 효과적인 고객 가치 분석을 가능하게 한다. 이탈 기준을 재정의하고, 등급에 따른 이탈률을 포함한 고객의 세분화된 군집별 특징을 기반으로 군집별 고객의 이탈 예측 모델을 제시한다. 이를 바탕으로 고객 들의 행동 패턴을 분석하여 고객 별 차별화된 마케팅 전략을 제공하기를 제안한다.

본 논문에서 제안하는 전체적인 모델은 고객 세분화, 이탈 예측, 고객 분석으로 구성되어 있다. 월별 RFM(Recency, Frequency, Monetary)을 데이터의 분야적 특성을 고려하여 변형한 새로운 RFM을 기준으로 고객을 세분화한다. 고객별 상이한 특성을 가장 잘 반영할 수 있는 최적 군집 수를 도출하고 K-Means Clustering을 통해 고객 세분화 과정을 거친다. 이후 클러스터별 이탈률과 월별 이탈률의 변동을 확인하여 각 클러스터링에 등급을 부여한다. 앞서 제시한 새로운 RFM 지표의 시계열적인 특성을 비롯한 고객 특성을 반영하여 고개의 이탈을 예측한다. EDA를 비롯하여, 표준화를 통한 카테고리별 비교와 TF-IDF방법을 활용한 상대적인 중요도로 군집별 특징적인 아이템을 살펴봄으로써 군집 별 특징을 도출해낸다. 이를 통해 군집별 고객 유지지 및 이탈 방지를 위한 차별화된 마케팅 전략을 제시함으로써 성공적인 고객 관계 관리(CRM)을 하는데 도움을 주고자 한다.



[그림 2.1-1] 전체 프로세스 구성

2.2 고객 세분화

기존 고객의 구매 패턴 파악하기 위해 RFM 방법론을 차용하여 고객 세분화를 수행한다. 하지만 본 연구는 월별로 변하는 패턴을 파악하는 것이 목적이기 때문에 고객이 얼마나 최근에 방문했는 지를 나타내는 지표인 기존의 R(recency)을 고객의 월별 평균 방문 주기로 변형한다. 또한 고객이 상품 구매하는 패턴과 제휴사를 이용하는 패턴이 다르다는 것을 반영하기 위해 두 구매 이력을 분리하여 지표를 설정한다. 따라서 변형된 RFM 으로 월별 고객 세분화를 진행한다.

2.2.1변형된 RFM

기존의 RFM 고객 세분화 지표는 아래와 같다.

- R(recency) : 고객의 마지막으로 구매한 시점으로부터 경과한 기간
  해당 기간이 짧을수록 고객가치가 더 높다고 판단할 수 있다. (1 년 전 방문한 고객보다 일주일 전 방문한 고객이 기업 입장에서 더 가치 있는 고객이다.)

- F(frequency) : 일정 기간 동안 고객의 구매 횟수
  해당 횟수가 높을수록 고객가치가 더 높다고 판단할 수 있다 . (기업의 서비스를 더 자주 이용하는 고객이기 때문이다.)
- M(monetary) : 일정 기간 동안 고객의 총 구매 금액
  해당 금액이 높을수록 고객가치가 더 높다고 판단할 수 있다. (기업에게 더 큰 이익을 가져다주기 때문이다.)

하지만 기존의 RFM 의 R(recency) 지표는 월별로 지속적으로 갱신되는 고객의 등급을 파악하기에 적합하지 않다. 예를 들어 두 고객의 1 월 구매횟수와 구매금액이 동일하다고 가정해보겠다. 만약 A 고객의 1 월 마지막 방문이 1 월 15 일, B 고객이 1 월 27 일이며 2 월에는 동시에 3 일에 방문했다고 가정한다면 A 고객의 1 월 R(recency)가 짧기 때문에 고객가치가 더 높다고 판단하는 것은 옳지 않다. 그러므로 고객의 최신성을 반영하는 R(recency) 지표 대신 고객이 월별로 같은 기간 동안 얼마나 빈번하게 방문하는 지를 반영한 평균방문주기(Avg_Visit) 지표를 제안한다.

평균 방문 주기는 고객이 한달 기준으로 처음 방문한 날로부터 마지막 방문일까지의 기간 동안 총 몇 번 방문했는지를 반영한다. 예를 들어 고객이 1 월 3 일부터 27 일 사이에 총 4 번 방문했다면, 고객의 평균 방문 주기는 8 일이 된다. 따라서 위의 값을 구하는 식은 아래와 같이 정의할 수 있다.

$$Avg\_Visit = \frac{\left(\textit{고객의 월별 마지막 방문일} - \textit{첫방문일} \right)}{\textit{월별 총 방문 횟수} - 1}$$

마지막으로 식품, 의류와 같은 상품 판매와 영화관, 식당과 같이 다양한 제휴 서비스를 제공하는 복합 쇼핑몰임을 고려하여, 두 서비스에 대한 고객의 구매 패턴이 다르다는 가정하에 두 경우를 분리하여 세분화를 진행하였다. 따라서 변형된 RFM 고객 세분화 지표는 아래와 같으며, 이후 해당 지표를 Avg_VFM 이라고 명명한다.

- P_Avg_Visit : 고객의 월별 평균 상품 구매 주기
  해당 주기가 짧을수록 고객가치가 높다고 판단할 수 있다. (평균 방문 주기가 3 일, 8 일인 고객을 비교할 때 3 일인 고객이 해당 서비스를 더욱 빈번하게 방문한다는 것을 알 수 있다.)
- P_Frequency : 고객의 월별 상품 구매 건수
  해당 건수가 많을수록 고객가치가 높다고 판단할 수 있다. (두 고객의 평균 방문 주기가 같을 때, 상품 구매 건수가 높은 고객이 더 다양하고 많은 물품을 구매했다는 것을 알 수 있다.)
- P_Money : 고객의 월별 총 상품 구매 금액
- A_Avg_Visit : 고객의 월별 평균 제휴사 방문 주기
  해당 주기가 짧을수록 고객이 자주 방문한다는 것을 알 수 있다.

- A_Frequency : 고객의 월별 제휴사 방문 횟수
  해당 횟수가 많을수록 고객이 제휴사 서비스를 더욱 다양하게 이용하고 있다는 것을 알 수 있다.
- A_Money : 고객의 월별 총 제휴사 이용 금액

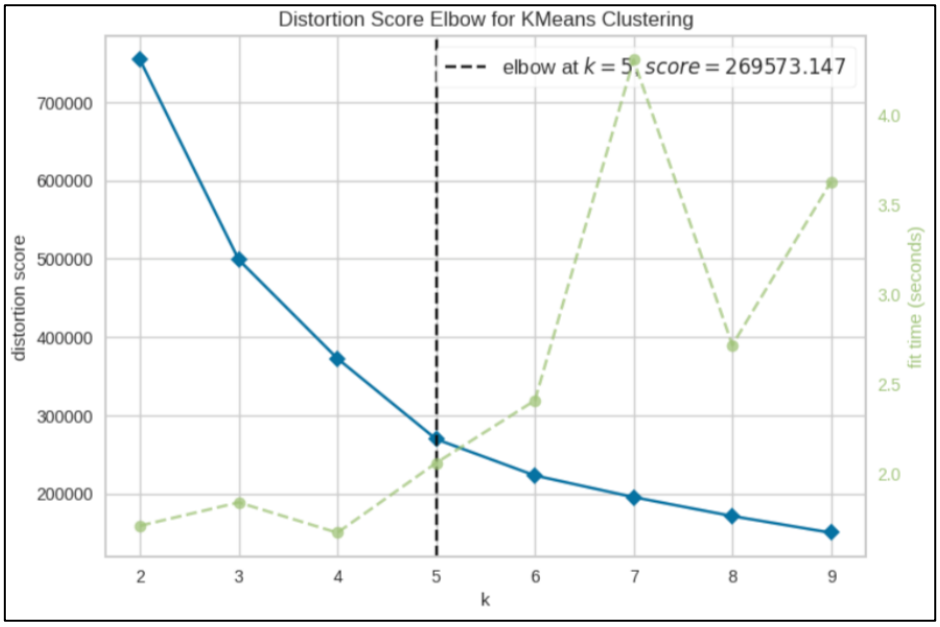
이로써 본 연구는 기존의 RFM을 변형하여, 월별로 고객의 방문주기를 반영하면서 서비스 종류에 따른 구매 패턴을 더욱 세분화여 살펴볼 수 있다.

2.2.2 K-Means Clustering 고객 세분화

기존의 전통적인 RFM 고객 세분화에서는 통계적으로 균등하게 집단을 나누어 점수를 부여한다. 하지만 이런 경우 각 군집에서의 특징이 두드러지지 않을 수 있다. 따라서 구매 패턴이 유사한 고객들로 군집을 형성해줄 수 있는 K-Means Clustering 기법을 사용하여 고객 세분화를 수행한다.

2.2.2.1 Elbow Method 최적 군집 개수 결정

1 년 동안 전체 방문 횟수가 5 회 미만인 신규 고객을 제외하고, 기존 고객들의 데이터를 2.2.1 에서 새롭게 정의된 지표를 사용하여 월별로 재구성한다. 그 후 다양한 군집수로 클러스터링 한 후 Elbow Method 를 사용하여 최적의 군집 수를 결정한다. [그림 2.2-1]을 통해 K 가 5 일 때 클러스터간 거리의 합이 급격하게 떨어지는 것을 확인할 수 있다.



[그림 2.2-1] 최적 군집 개수 결정

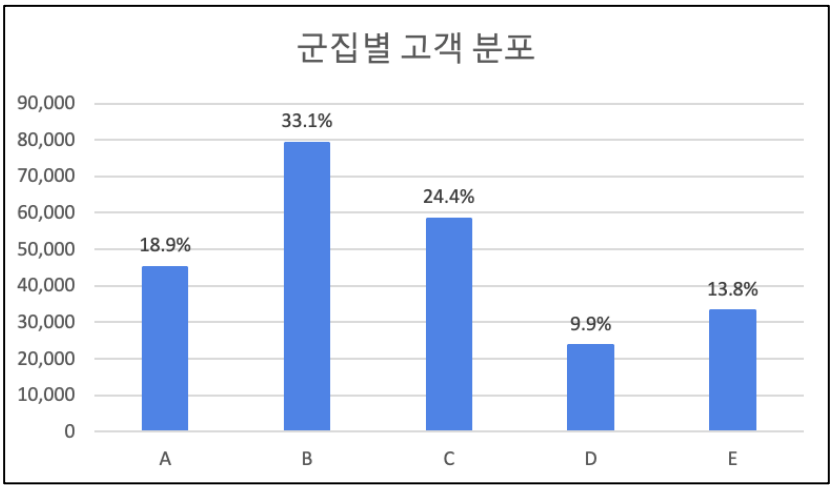
2.2.2.2 고객 세분화 군집 결과

위의 과정에서 구한 최적 개수인 K=5 로 클러스터링을 수행한다. 부여된 클러스터를 기준으로 각 데이터들의 평균을 계산한 후 비교하여 군집 별 특징을 살펴본다. [표 2.2-1]는 각 군집에 대한 결과를 정리한 표이다.

Cluster	P_Avg_Visit	P_Frequency	P_Money	A_Avg_Visit	A_Frequency	A_Money
A	9.02	9.34	701011.57	6.35	4.09	85237.30
B	8.25	8.54	671439.92	28.21	1.48	31902.57
C	5.19	8.08	605304.74	30.00	0.00	0.00
D	25.13	1.84	150329.04	29.99	0.1	1123.8
E	30.00	0.05	243.19	22.21	2.01	34558.43

Avg_Visit : 기간(일), Frequency : 구매건수(번), Money : 총 구매금액(원)

[표 2.2-1] 군집 별 평균 RFM



[그림 2.2-2]

각 군집의 고객 수 분포는 위의 그래프와 같다. 예시로 B군집을 살펴보면, 상품 구매는 짧은 방문 주기로 빈번하게 구매하지만 제휴사는 긴 방문 주기로 적게 이용하는 것을 알 수 있다. 따라서 상품 구매 패턴과 제휴사 이용 패턴을 복합적으로 고려하여 각 클러스터의 해석하면 아래와 같다.

- A : 상품과 제휴사를 모두 많이 이용하며 사용 금액이 높은 군집
- B : 상품과 제휴사를 모두 이용하지만 A 군집에 비해 제휴사는 덜 이용하는 군집
- C : 상품 구매만 높은 빈도로 이용하는 군집
- D : 상품만 가끔 구매하는 군집
- E : 제휴사만 가끔 이용하는 군집

2.3 이탈 등급 분류

2.3.1 이탈 정의

통신사 분야나 금융 업계에서는 고객 이탈은 기존 고객이 계약하고 있던 통신사 변경이나 계좌, 보험, 카드를 해지하는 시점으로 이탈을 정의하므로 그 기준이 명확하다. 하지만 본 연구와 같은 특정 기업의 제휴사 이용 고객의 경우는 이탈을 판단할 기준이 명확하게 존재할 수 없다. 이에 유사 연구들은 실질적으로 이탈의 기준을 실무자의 주관에 따라 정한 뒤 연구를 진행하는 방식을 취하고 있다.

따라서 본 연구에서는 이탈 유무에 대한 기준을 한달을 기점으로 구매 이력의 유무를 기준으로 정의하였으며, 세부적으로 고객의 첫 구매 월 시점을 기준으로 이후 특정 월에 구매 이력이 없는 경우, 그 한달의 기간 동안 고객이 이탈했다고 정의했다. 예를 들어, A 고객의 첫 구매가 1 월인 고객이 6 월에 방문하고 7 월에 방문하지 않았다면, 고객을 7 월에 기점으로 이탈 고객으로 정의하였다.

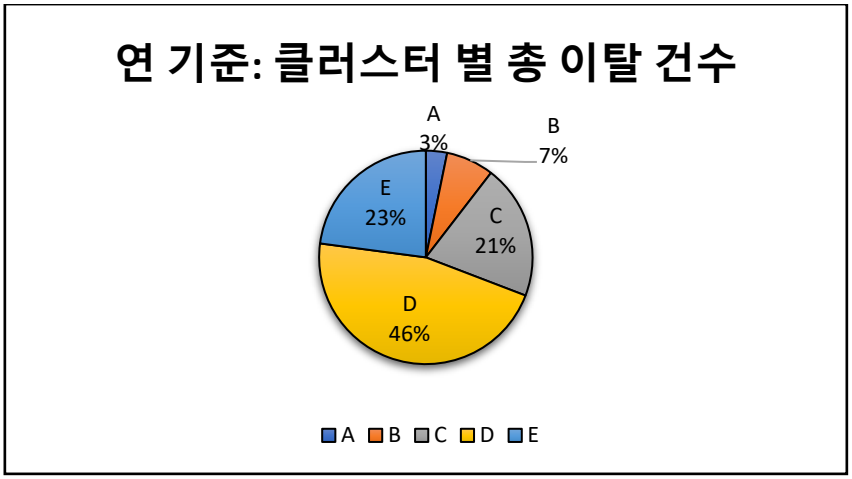
2.3.2 클러스터별 이탈 확인

위와 같이 고객의 이탈을 정의한 뒤 5 개의 고객 클러스터별 이탈률을 산출하였다. 이때 고객별 이탈률을 구할 경우 두가지를 고려하여야 한다. 먼저, 고객의 이탈은 개개인의 첫 구매 시점을 기준으로 전체 이탈의 수를 집계하여야 한다. 또한, 1 월의 경우에는 이전 년도 12 월의 구매 데이터가 없으므로 클러스터별 이탈률을 계산할 수 없다.

2.3.2.1 이탈 건수

① 연 기준: 클러스터별 이탈 건수의 크기

1 년 동안 일어난 클러스터별 이탈 건수의 크기는 다음과 같다.

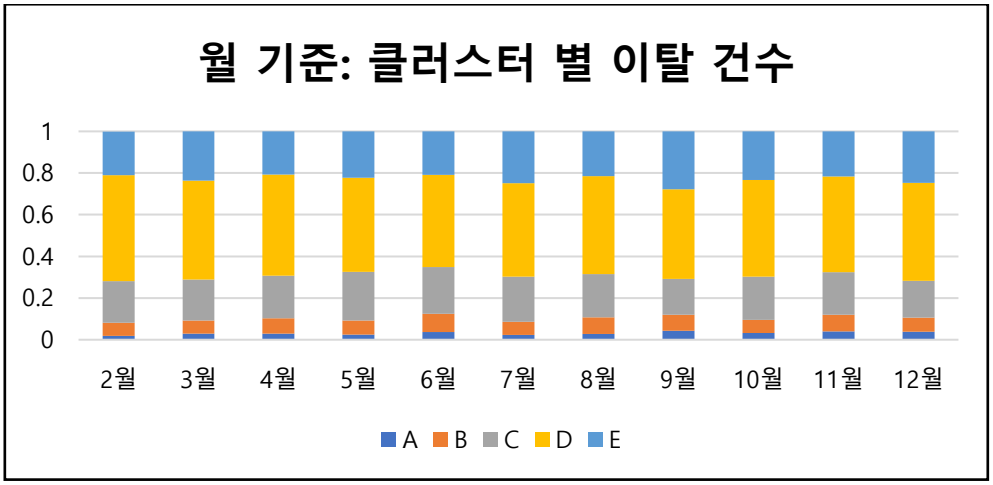


[그림 2.3-1]

12개월 동안 일어난 이탈의 수는 총 44,381건이다. 이중 이탈 발생 이전 달에 구매 이력이 있는 경우는 44,381건에서 이전달에 구매이력이 없는 경우를 제외하여 구하였고, 총 27,000건으로 나타났다. 최종 27,000건의 이탈에서 이탈 시점 이전 달의 고객의 클러스터는 각 A인 경우 810건, B인 경우 1,890건, C인 경우 5,670건, D인 경우 12,420건, E인 경우 6,210건이다. 이를 통해 구한 전체 27,000 이탈건수에 대한 각 클러스터의 차지 크기는 위 그림과 같다.

② 월 기준: 이탈에 대한 클러스터별 차지 크기

전체 월별 이탈에 대한 클러스터의 비율은 다음과 같다.



[그림 2.3-2]

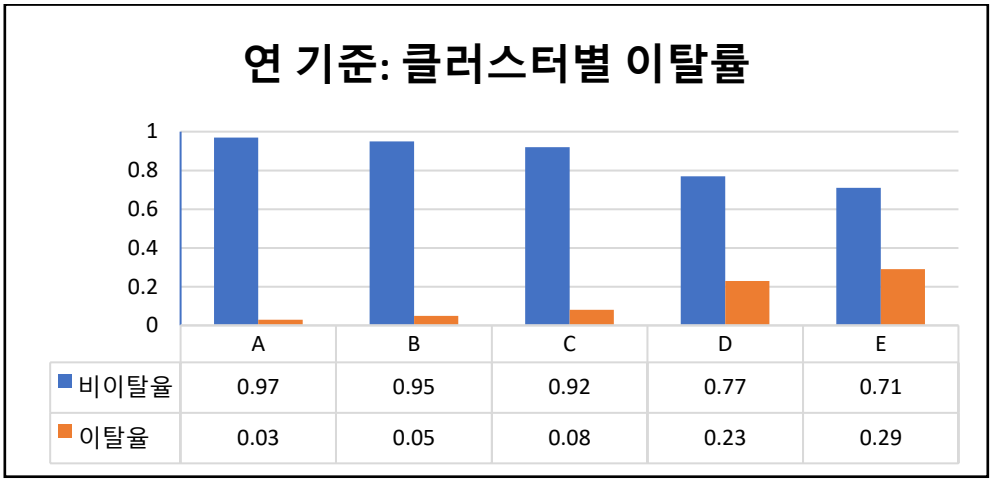
월별로 월의 이탈건수를 구하고, 각 이탈건수에 대해 이전 달이 어떤 클러스터였는지를 카운트하였다. 그 결과는 위 그림과 같다. 각 월별 클러스터별 이탈 건수는 일반적으로 D, E, C, B, A 순으로 그 건수가 많으나, 5 월과 6 월은 E 클러스터에서 C 보다 이탈 건수가 많이 나타난다. 하지만 본래 E 와 C 의 이탈건수 차이가 미비하기 때문에 유의한 변화라고 보기 어렵다. 결론적으로 월을 기준으로 각 클러스터별 이탈건수가 비슷하게 나타나며 이 경향성이 1 년동안 유지되는 것으로 판단된다.

하지만 위 결과를 통해 각 클러스터별 이탈비율을 비교할 수는 없다. 이는 군집화를 통해 생성된 군집의 크기는 서로 상이하기 때문이다. 따라서 각 클러스터의 수를 고려한 상태에서의 클러스터의 이탈률을 구하고 비교하는 과정이 필요하다.

2.3.2.2 이탈 비율

① 비율연 기준: 클러스터별 이탈률

각 클러스터별로 크기를 고려하여 구한 클러스터별 이탈률은 다음과 같다.

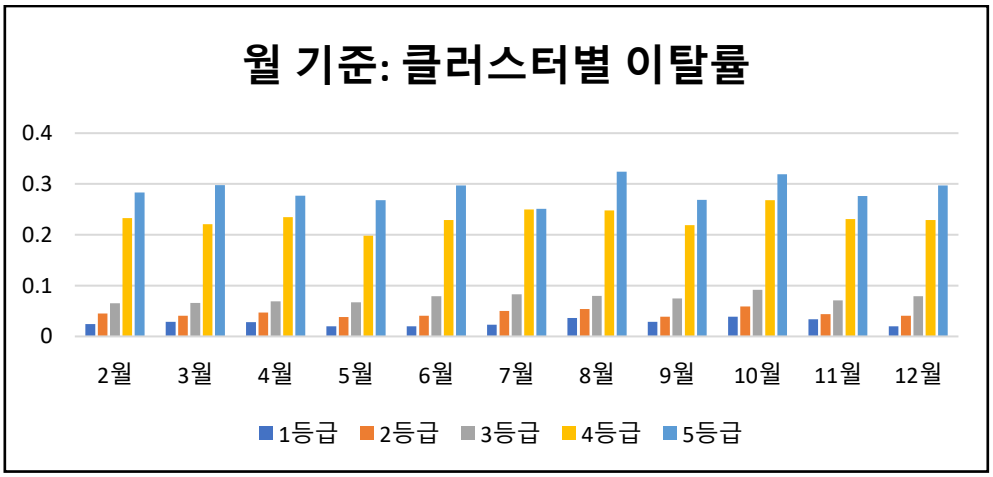


[그림 2.3-3]

주황색 막대가 이탈률을 의미하며, 기존 D, E, C 순으로 이탈의 건수는 많았던 것에 비해 클러스터의 인원 크기를 고려한 각 클러스터의 이탈률로 본다면 C, D, E 순으로 이탈률이 높아짐을 확인할 수 있다. 결과적으로 클러스터 별 이탈률은 A, B, C, D, E 순으로 높게 나타난다. 이 클러스터별 이탈률 결과를 기반으로 이탈률이 낮은 A 부터 E 순으로 각 클러스터에 1 등급~5 등급으로 등급을 부여하였다.

② 월 기준: 클러스터별 이탈률

월 기준 클러스터별 이탈률은 다음과 같다.



[그림 2.3-4]

클러스터별 이탈률을 월 기준으로 보더라도 각 클러스터별 이탈률이 위에서 부여한 등급 순으로 높음을 확인할 수 있고, 그 순위가 각 월별로 동일하게 유지됨을 확인할 수 있다.

2.4 이탈 예측 모델

2.4.1 필요성

본래 클러스터 중 이탈률이 높은 클러스터를 대상으로 이탈을 방지하는 마케팅을 실시하려 하였으나, 각 클러스터 자체의 인원수가 많고, 이탈률 또한 비이탈률이 더 높아 클러스터 자체가 이탈 방지 마케팅에 대해 유의미한 고객만을 제시해 준다고 보기 어렵다고 판단하였다. 또한 이탈 관리의 경우 이탈에 대한 손해정도 측면>에서는 충성고객의 이탈을 관리하는 것이 더 효율적이나, 이탈 자체는 일반 고객에서 더 많이 일어나므로 그 타겟을 한정 짓는 것은 비효율적이다. 따라서 특정 클러스터만을 지정하여 이탈을 관리하는 것보다 모든 클러스터를 대상으로 이탈 예측을 적용해야 한다고 판단하였다.

각 클러스터별 이탈률은 가장 작은 경우가 1 등급 2%, 큰 경우가 5 등급 29%이다. 5 등급의 이탈률은 1 등급의 14.5 배로 클러스터별 이탈률은 유의미하게 차이가 난다. 또한 1 월달에 등급이 1 등급이였다가 2 월달에 4 등급으로 등급이 하락하는 것과 같이, 2 개월 동안의 등급의 하락이 존재하고 하락 다음달에 이탈이 일어나는 경우를 살펴보았을 때, 등급의 하락 이후 이탈이 일어나는 경우은 모든 이탈에 대해 10%정도로 나타났다. 이에 고객의 Avg_VFM 데이터 자체와 Avg_VFM 데이터의 시계열적 변화를 통해 고객의 이탈 예측이 유의할 것으로 판단하여, 시계열적 Avg_VFM 데이터 구성을 통한 고객의 이탈 예측 모델링을 실시하였다.

2.4.2 데이터 구축 및 모델링

2.4.2.1 데이터 구성

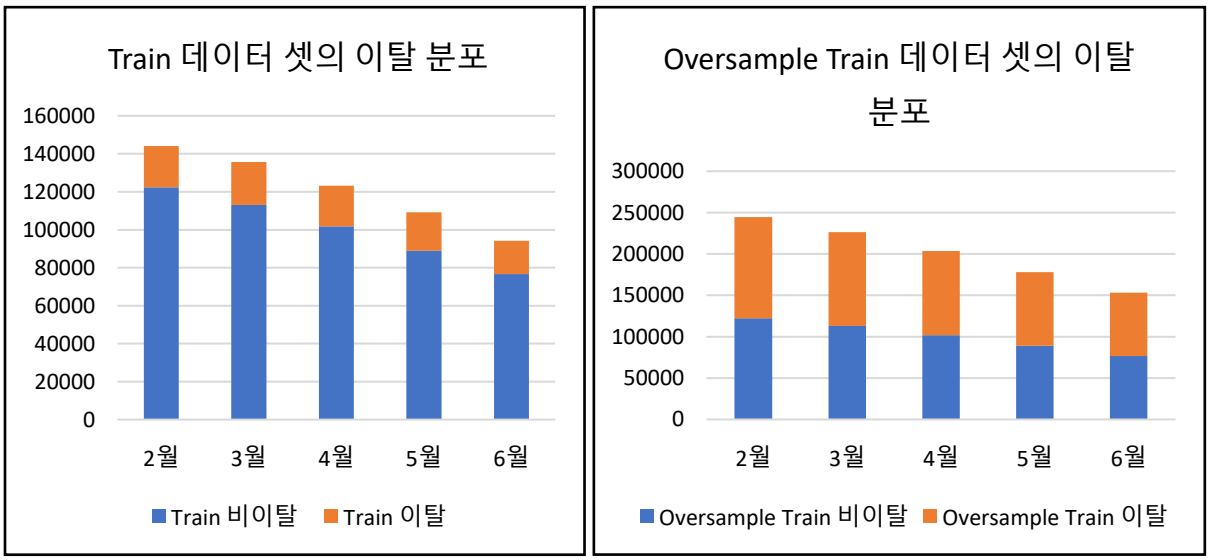
고객별 Avg_VFM 의 시계열적 변화를 데이터로 넣기 위해 데이터의 열을 고객의 월별 Avg_VFM 데이터로 구성하였다. 만약 모델이 6 개월 동안의 고객 Avg_VFM 데이터를 기반으로 다음달 이탈을 예측한다면, 데이터프레임의 열은 고객의 6 개월 이전 달의 Avg_VFM 값부터 순차적으로 근접한 달의 Avg_VFM 값을 가지게 된다. 또한 각 열 이름을 '6 개월 전', '5 개월 전'~ '2 개월 전', '1 개월 전'으로 지정하여, '6 개월 전' 열에 이탈 평가 시점을 기준으로 1 월, 4 월, 6 월 등 다양한 월이 들어갈 수 있도록 하였다. 위 예시대로 형성된 6 개월 기준 예시 데이터는 다음 그림과 같다.

고객 번호	6개월 전	5개월 전	~	1개월 전	이탈 여부
A_1	A_1의 1월 Avg_VFM	A_1의 2월 Avg_VFM	~	A_1의 6월 Avg_VFM	A_1의 7월 이탈 여부
A_1	A_1의 4월 Avg_VFM	A_1의 5월 Avg_VFM	~	A_1의 9월 Avg_VFM	A_1의 10월 이탈 여부

[표 2.4-1] 6 개월 반영 데이터 셋 예시

각 이탈여부를 판정하는 기준이 7 월 10 월인 경우 각 '6 개월 전' 열에 특정 고객의 1 월 데이터, 4 월 데이터가 들어가게 된다. 이때 고객의 특정 월에 이탈로 데이터가 없는 경우는 각 값이 클수록 의미상 좋지 않은 Avg_Visit 변수에는 30 을, 값이 작을수록 의미 상 좋지 않은 Frequency 와 Monetary 변수에는 0 을 넣어 NaN 값을 처리하였다. 또한 각 월별 Avg_VFM 데이터는 6 열로 구성되므로 최종 데이터에서 Avg_VFM 열은 6x6 인 36 열로 형성된다.

입력 데이터의 전처리는 로그변환, standard 스케일링의 과정을 따랐으며, 추가적으로 각 월을 기준으로 이탈의 여부를 판정한 경우 이탈과 비이탈의 데이터 불균형이 크게 존재하여, 오버샘플링 과정을 학습 데이터에 적용하였다.



[그림 2.4-1] Train 데이터 셋의 Oversample 전후 이탈 분포

기존 데이터 셋의 경우 전체 셋 기준 15~18%가 이탈 고객으로 데이터의 불균형이 존재하였으며, 이에 RandomOverSample 방식을 사용하였다.

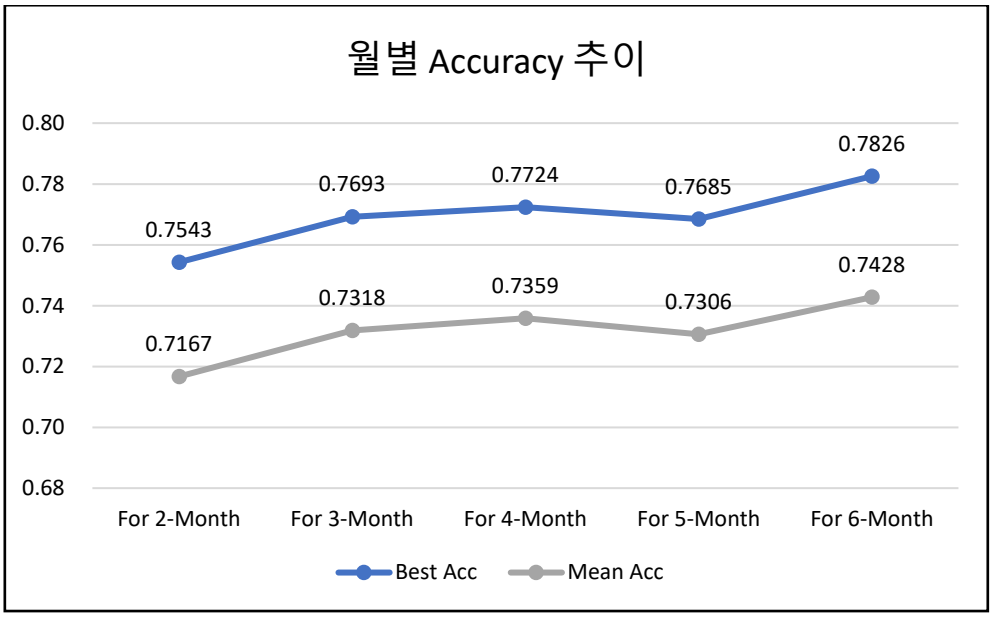
2.4.2.2 월 길이 별 모델 성능 비교

시계열 적으로 Avg_VFM 데이터를 넣었을 때의 적절한 기준으로 파악하기 위해 각 2,3,4,5,6 개월을 기준으로 데이터를 열로 생성하여, Validation Set 에 대한 각 모델의 평가지표를 비교하였다. 이 때 모델 별 하이퍼파라미터는 통일하였고, 평가지표의 경우 Accuracy, Precision, Recall, Auc 값을 사용하였다. 통일 파라미터는 [부록 1]와 같다.

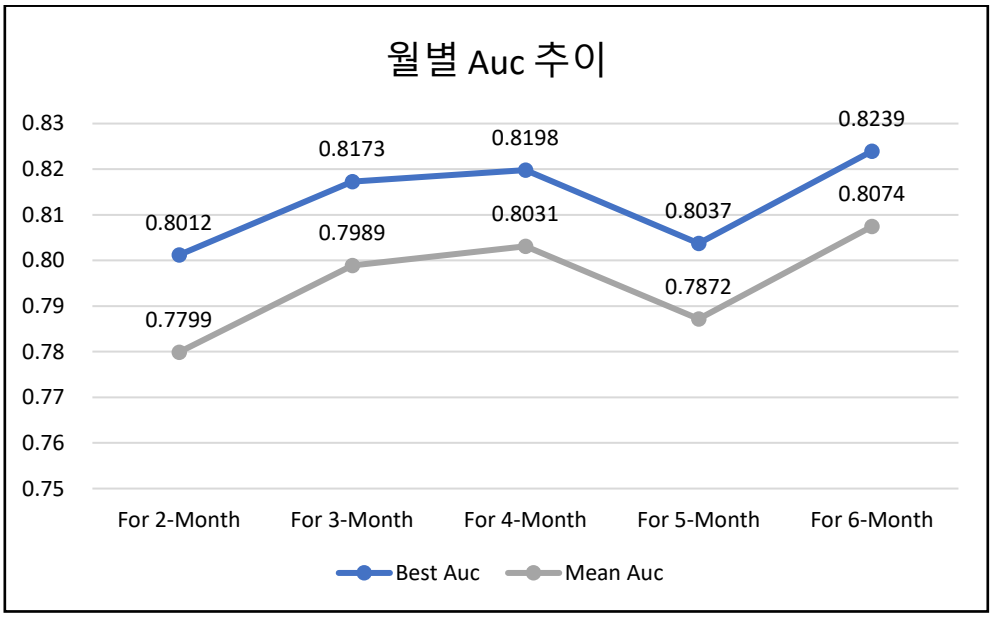
성능지표_Validation		**Accuracy**	**Precision**	**Recall**	**Auc**
2개월	Logistic	0.694	0.305	0.789	0.798
	XGB	0.701	0.311	0.784	**0.801**
	RandomForest	**0.754**	0.338	0.638	0.794
	CatBoost	0.737	0.321	0.644	0.773
	GBM	0.713	0.321	0.756	0.797
	K-NN	0.701	0.270	0.565	0.716
3개월	Logistic	0.710	0.344	0.806	**0.815**
	XGB	0.719	0.351	0.796	0.817
	RandomForest	**0.769**	0.389	0.660	0.813
	CatBoost	0.759	0.369	0.621	0.784
	GBM	0.730	0.357	0.764	0.813
	K-NN	0.704	0.314	0.648	0.751
4개월	Logistic	0.710	0.359	0.805	0.816
	XGB	0.725	0.371	0.789	**0.820**
	RandomForest	**0.772**	0.410	0.644	0.818
	CatBoost	0.769	0.397	0.575	0.783
	GBM	0.736	0.378	0.754	0.816
	K-NN	0.701	0.337	0.700	0.766
5개월	Logistic	0.707	0.368	0.788	0.803
	XGB	0.721	0.378	0.763	**0.804**
	RandomForest	**0.769**	0.420	0.617	0.802
	CatBoost	0.764	0.402	0.540	0.764
	GBM	0.732	0.387	0.731	0.800
	K-NN	0.691	0.341	0.695	0.750
6개월	Logistic	0.716	0.374	0.802	0.819
	XGB	0.735	0.391	0.779	**0.824**
	RandomForest	**0.783**	0.436	0.619	0.822
	CatBoost	0.777	0.417	0.525	0.787
	GBM	0.747	0.399	0.736	0.819
	K-NN	0.698	0.349	0.739	0.774

[표 2.4-2] 월 길이와 모델 별 성능지표

Accuracy, Auc 기준 월 기준 별 평가지표의 최대, 평균값은 아래 그림과 같다



[그림 2.4-2] 월별 성능지표 통계량 비교_Accuracy

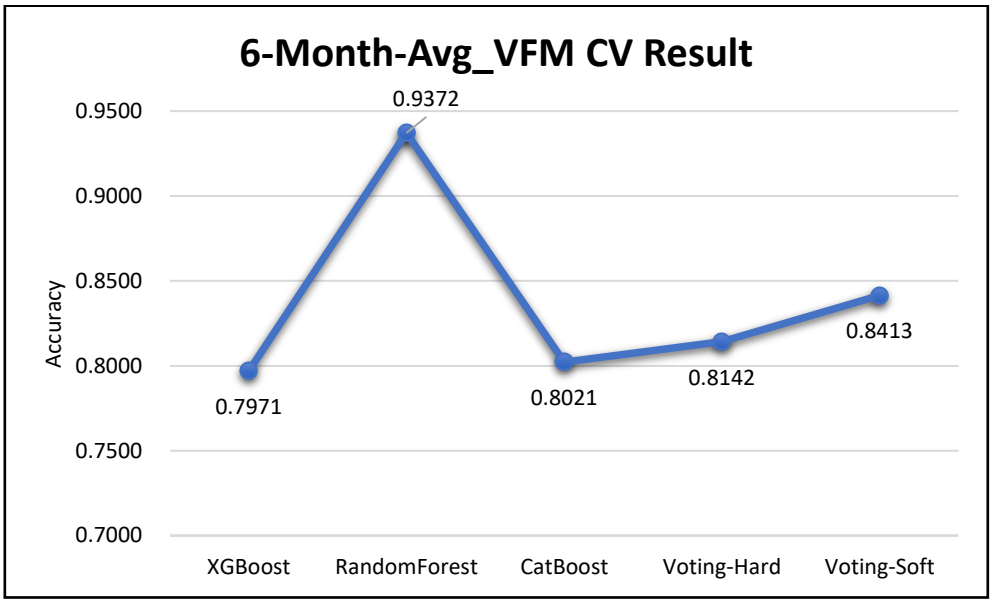


[그림 2.4-3] 월별 성능지표 통계량 비교_Auc

성능 지표의 통계값을 확인 한 결과 데이터의 포함 월 기간이 증가할수록 성능지표가 향상됨을 확인할 수 있으나, 5 개월 단위의 경우는 오히려 지표가 하락함을 확인할 수 있다. 또한 월이 길어질 경우 그 학습 데이터의 절대적 양이 줄어들며, 데이터의 설명력과 학습의 어려운 정도가 높아진다. 결론적으로 개월 수가 길되 5 개월을 적용하는 것은 부적절하다고 판단하여 최종 6 개월의 길이로 데이터 월 길이를 지정하였다

2.4.2.3 6개월 기준 모델 별 성능 비교

모델링 결과에서 정확도가 높은 기준으로 RandomForest 와 CatBoost 모델을을 Auc 가 높은 기준으로 XGBoost 모델을 후보로 선정하여, 각 모델에 대한 K-Fold Cross Validation 을 적용하였다. 추가적으로 각 최고 모델을 기반으로 Voting 모델을 적용하였다. 결과는 다음과 같다.



[그림 2.4-4] 6-Month-Avg_VFM CV Accuracy 모델 별 결과

2.4.2 비교군 모델링

① 1 개월 데이터

데이터가 비시계열적으로 존재하는 경우와 결과를 대조하기 위하여, 고객의 이전 달의 Avg_VFM 데이터를 기반으로 다음달의 이탈을 예측하는 모델을 생성하였다. 그 성능은 다음과 같다.

모델		AVG_Accuracy
1 개월	XGB	0.6601
	RandomForest	**0.8811**
	CatBoost	0.6603

[표 2.4-3] 비시계열적 Avg_VFM 데이터 기반 이탈 예측 결과

② 고객 특성 변수 추가

기존 Avg_VFM 6 개월 데이터에 고객 특성 변수인 성별, 나이대, 기업 결제 수단 사용 무유를 변수로 추가하였다. 각 모델의 성능은 다음과 같다.

모델		AVG_Accuracy
6 개월	XGB	0.7972
	RandomForest	**0.9378**
	CatBoost	0.8031

[표 2.4-4] 고객 특성 변수 추가 데이터 기반 이탈 예측 결과

③ Recenvy 변수 추가

본래 클러스터상 군집화를 위해 Avg_Visit 을 사용하였으나, 다음달의 고객의 방문여부를 예측하는 문제에서는 고객의 가장 최근 방문일로부터 얼마가 지났는지가 유의하게 영향을 줄 수 있을 것이라고 판단되어 Recency 변수를 변수에 추가해 주었고, 그 결과는 다음과 같다.

모델		AVG_Accuracy
6 개월	XGB	0.7142
	RandomForest	**0.8601**
	CatBoost	0.7062

[표 2.4-5] Recency 변수 추가 데이터 기반 이탈 예측 결과

④ 고객 특성 변수와 Recency 변수 추가

②에서의 고객특성 변수와 Recency 를 모두 넣은 모델의 결과는 다음과 같다.

모델		AVG_Accuracy
6 개월	XGB	0.7105
	RandomForest	**0.8599**
	CatBoost	0.7042

[표 2.4-6] 고객 특성 변수와 Recency 추가 이탈 예측 결과

⑤ 일반 이탈 예측 모델

일반 모델을 구축함에 있어, 고객의 한달 전 각 제휴사(제휴사 코드 기준, 총 12 가지)의 방문 횟수와, ②에서 사용한 고객 특성 변수와 구매금액을 사용하였다, 결과는 다음과 같다.

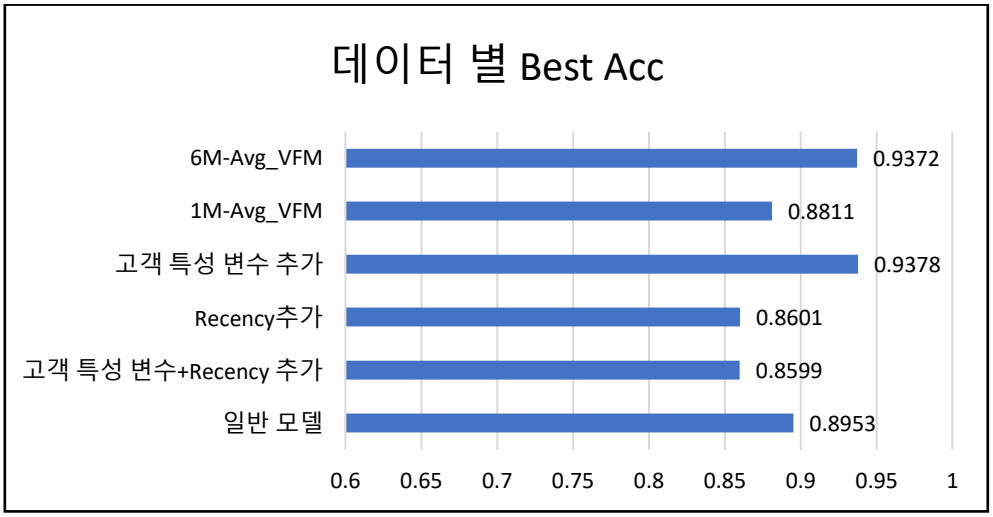
모델		AVG_Accuracy
1 개월	XGB	0.7371
	RandomForest	**0.8953**
	CatBoost	0.7391

[표 2.4-7] 일반 이탈 예측 모델 예측 결과

2.4.4 결과 비교 및 분석

2.4.4.1 최적 데이터 및 모델 선정

위 기존 모델과 비교 모델들의 최종 Cross-Validation 값의 비교 결과는 다음과 같다.



[그림 2.4-5] 데이터 구성별 Best Accuracy 결과

위 6 가지 데이터 중 6 개월의 Avg_VFM 데이터에 고객 특성 변수를 추가한 데이터를 기반으로 모델링 했을 경우 Best Model 의 Cross-Validation 결과가 가장 높았다. 이 때 Best Model 은 RandomForest 모델이며, 그 평균 정확도는 0.9378 로 측정되었다.

2.4.4.2 모델 선정 및 test 결과

최종 최적으로 선정된 데이터와 Best 모델에 대해 하이퍼파라미터 튜닝을 적용하였고, 그 결과 최종 기존 default 하이퍼파라미터에서 가장 최대의 정확도가 측정되었다. 이때 Grid 범위와 선정 Grid 는 [부록 2]와 같다.

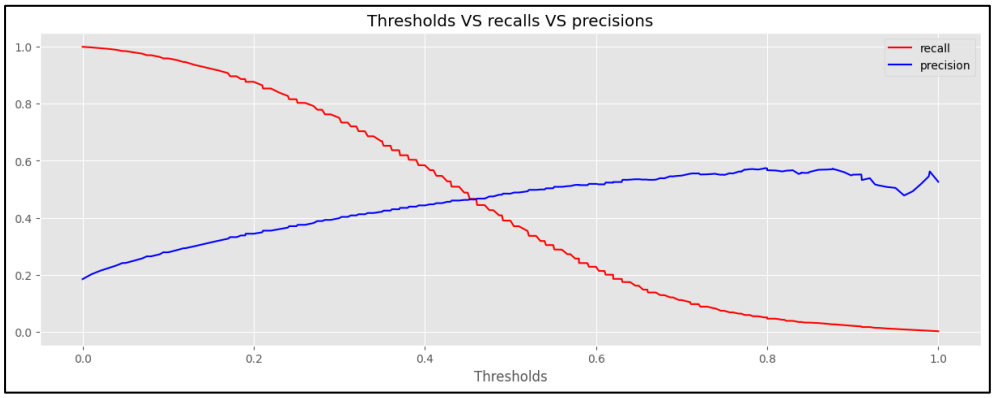
최적 하이퍼파라미터를 적용한 모델의 Test Set 에 대한 Confusion matrix 와 각 성능 지표 값은 다음과 같다.

RandomForest Classifier		**예측**	
		Predict_이탈	**Predict_비이탈**
실제	**이탈**	2016	3421
	비이탈	2162	21851
Accuracy		81.04%	
Precision		48.25%	
Recall		37.08%	
Auc		81.84%	

[표 2.4-8] 최종 모델 Confusion Matrix

모델의 Accuracy, Auc 를 기반으로 80%이상의 정확도를 보이며, Threshold 값 0.5 를 기준으로 Recall 값은 이전 Validation 에서의 결과들에 비해 하락하였으나, Precision 값이 증가한 것을 볼 수 있다.

모델의 Threshold 값 변화에 따른 Recall 과 Precision 변화는 다음과 같다.



[그림 2.4-6] Threshold 값 변화에 따른 재현율과 정밀도 변화

기존 모델에서 Threshold 이 0.2 인 경우 Precision 는 35%, Recall 은 86.7% 로 측정된다. 만약 위 모델을 기반으로 이탈 고객을 타겟 하려는 경우, 각 마케팅의 비용과 마케팅 대상별 기대효과를 비교하여, 임계값을 조정하는 과정이 필요할 것으로 보인다.

최종 모델을 기반으로 이탈이라고 예측된 인원 2016 명에 대해, 각 한 고객을 기준으로 6 개월 동안의 월 별 등급을 알 수 있으므로, 6 개월 이전부터 한달 전까지 이탈이라고 예측된 인원들의 등급을 카운트한 결과는 다음과 같다. (특정 월에 구매 이력이 없는 경우는 제외한 값이다.)

등급	6 개월 전	5 개월 전	4 개월 전	3 개월 전	2 개월 전	1 개월 전
1 등급	10	19	4	5	1	0
2 등급	45	32	35	18	6	2
3 등급	357	315	257	153	56	8
4 등급	509	497	466	413	265	214
5 등급	382	427	435	443	358	249

[표 2.4-9] 이탈 예측 고객 등급 분포

전반적으로 1~5 등급에 대해 순위가 낮은 4 등급, 5 등급인원에 대해 이탈을 예측하나, 수가 적고 이탈의 비율도 2%밖에 되지 않는 1~2 등급에 대해서도 이탈 인원을 예측한다.

2.4.4.3모델 비교·평가 및 의의

1) 시계열적 데이터 구성

기존 1 개월 동안의 고객의 Avg_VFM 을 기반으로 고객의 다음달 예측을 진행한 결과는 6 개월 동안의 Avg_VFM 변수를 고려한 모델 보다 정확도가 5.5%낮게 측정되었다. 이는 각 고객의 Avg_VFM 데이터 뿐만 아니라 Avg_VFM 상의 그 변동의 경향성이 예측 확률을 높여준다는 것을 의미한다.

2) Avg_Visit 을 통한 Recency 변수의 대체

[표 2.4-6]의 결과를 보면 Avg_VFM 데이터에 Recency 변수를 추가한 경우 그 정확도는 8%가량 감소한다. 이는 Frequency 를 변수로 넣은 데이터의 경우 본래 Avg_VFM 데이터 상 2~3 이였던 제휴사_Visit 변수의 Vif 값이 10 까지 증가하므로 (Vif 변화는 부록 3 참조) Frequency 와 Avg_Vitis 변수와의 다중공산성이 그 원인으로 파악된다. 이를 통해 Avg_Vitis 변수가 Recency 변수를 분류 모델에서 충분히 대체하고 있음을 확인할 수 있다.

3) 고객 특성 변수의 추가

클러스터링 기반으로 각 군집간 유의하게 차이가 보이는 변수인 성별, 나이대, 기업 결제 수단 사용 유무를 데이터에 추가하여 모델링하였을 경우 모델의 예측 정확도는 상승하였다. 이는 추가된 고객의 특성 변수와 Avg_VFM 변수 사이에 유의미한 상호작용이 존재하여 더욱 정확한 예측 결과를 도출함을 의미한다.

4) 이탈 예측 정도와 예측 대상 범위

이탈 예측 모델의 경우 최종 6 개월 Avg_VFM+고객특성변수 데이터 기반 RandomForest Classifier 모델이 최종 모델로 선정되었으며, Test Set 에 대해 정확도와 Auc 기준 81% Precision 기준 48%, Recall 기준 37%로 성능 지표가 측정되었다. 이는 기존 클러스터 중 이탈률이 높은 클러스터를 단순히 이탈 마케팅의 대상으로 타게팅하는 것보다 더 나은 이탈 예측 타겟을 동일 데이터와 모델을 기반으로 타게팅 해준다고 할 수 있다. 만약 마케팅 비용을 조금 감수하여, 이탈 임계값을 낮추고 마케팅 대상을 늘릴 경우, 임계값 0.2 기준 Precision 은 35%로 약 12%하락하나 Recall 값은 86.7%로 약 50% 상승한다. 이는 다음달 이탈 전체 이탈 인원 중 86.7%이 마케팅 대상에 포함됨을 의미한다.

위 [표 2.4-9]를 통해 모델이 단순히 변수의 값이 작거나, 좋지 못한 4 번,5 번 군집을 이탈로 예측하는 것이 아닌, 5 단계 등급에 대해 이탈 인원을 예측하고 있음을 확인할 수 있다. 이는 이탈 예측 모델이 Avg_VFM 의 시계열적 변화를 고려하고 있다는 또 하나의 증거이며, 최종 모델이 이탈 방지의 리턴 효과가 큰 충성고객부터 이탈 자체가 많이 일어나는 일반 고객까지의 모든 범주에 대해 이탈을 예측해주는 모델임을 의미한다.

5) 데이터의 간결성

고객등급관리에 있어 필수적으로 존재하는 데이터인 RFM 데이터에 단순한 고객 특성 변수만을 추가하여 고객의 이탈 예측 모델을 구축하므로, 모델의 설명력과 데이터 간결성 측면에서 그 의의가 있다.

2.5 군집 별 구매 아이템 특성 확인

상품 구매와 제휴사 이용 패턴을 반영한 5 개의 군집 간 고객들의 구매 아이템 차이가 존재한다면, 각 군집 간 고객의 특성 차이를 파악하는데 도움이 될 것이다. 또한 해당 아이템들을 활용하여 마케팅에 적용할 수 있을 것이다.

하지만 해당 데이터의 특성 상, 대부분의 고객들이 구매 주기가 짧고 가격이 저렴한 식품 카테고리를 보편적으로 많이 구매한다. 따라서 구매 횟수로 각 상품 카테고리를 비교할 경우 식품이 우세하게 되어 정확한 비교가 어렵다. 또한 구매 금액으로 고려할 경우 가전이나 가구와 같은 특정 아이템의 금액대가 매우 높아 각 군집 별로 다른 상품에 대한 특징을 도출하기 어렵다.

따라서 본 연구에서는 표준화를 통한 카테고리별 비교와 고객들이 보편적으로 구매하는 아이템의 값을 조절하여 상대적인 중요도로 군집 별 특징적인 아이템을 살펴보고자 한다.

2.5.1 Standardization

표준화는 데이터를 특정 범위로 변환하는 과정으로 데이터의 스케일을 조정하여 데이터 간의 상대적인 크기 차이를 보정하여 일관된 척도로 데이터를 비교하고 분석할 수 있도록 도와주는 방법론이다.

본 과정에서는 월별로 각 고객의 카테고리별 구매건수를 행렬로 구성한 데이터를 사용한다. 카테고리별 특성에 따라 다양한 구매건수의 단위 차이를 동일하게 맞추기 위해 열별(카테고리별)로 표준화(standardization)을 수행하고 변환된 결과를 비교한 표는 아래와 같다. (전체 행렬 중 일부만 추출하였다.)

고객	스낵류	잎채소	국산과일	이용권/입장권	스케이트/썰매	사무기기
A	2	3	1	0	0	0
B	0	2	1	0	0	0
C	1	0	0	1	0	0
D	0	1	0	0	1	0
E	2	1	0	0	0	1

[표 2.5-1] 표준화 적용 전

고객	스낵류	잎채소	국산과일	이용권/입장권	스케이트/썰매	사무기기
A	0.475	1.940	0.337	-0.046	-0.005	-0.005
B	-0.419	1.168	0.337	-0.046	-0.005	-0.005
C	0.028	-0.375	-0.406	5.088	-0.005	-0.005
D	-0.419	0.397	-0.406	-0.046	**190.320**	-0.005
E	0.475	0.397	-0.406	-0.046	-0.005	**70.277**

[표 2.5-2] 표준화 적용 후

두 표를 통해 보편적으로 많이 구매되며 구매건수가 상대적으로 큰 스낵류의 경우 정규분포 형태로 값이 조절되지만 "이용권/입장권", "스케이트/썰매", "사무기기"와 같이 구매가 많지 않은 아이템의 경우 지나치게 높은 값으로 설정된 것을 확인할 수 있다.

2.5.2 TF-IDF

텍스트마이닝의 기법 중 하나인 TF-IDF(Term Frequency-Inverse Document Frequency) 분석은 가중치가 부여된 단어를 통해 문서의 특징을 표현하여 두 문서 간 유사도를 비교하거나 문서의 핵심어를 추출하는 방법이다. 여러 문서 속에서 각 단어가 특정 문서 내에서 얼마나 큰 중요도를 가진 정보인지를 나타내는 통계적 수치를 보여준다. 큰 값을 가질 수록, 해당 문서 내에서 중요한 의미를 갖고 있는 것으로 해석할 수 있다.

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

TF(d,t)는 단어 t가 문서 d에서 등장하는 횟수를 말하며 출현 횟수가 많을수록 중요한 단어로 판단한다. IDF(t,D)는 다른 문서에서 흔하게 나타나는 단어일수록 가중치를 낮추기 위한 값이며 전체 문서 D의 수를 단어 t가 포함된 문서의 수 + 1의 값으로 나눈 다음 로그를 취하는 것이 일반적이다. 분모에 1을 더하는 이유는 t가 포함된 문서의 수가 0일 때 분모가 0이 되는 것을 방지하기 위함이며 TF-IDF는 TF(d,t)와 IDF(t,D)의 곱으로 나타낸다. 본 논문에서 단어 t는 상품 카테고리를, 문서 d는 각 고객을 의미한다. 즉, TF-IDF가 높은 단어를 통해 모든 고객에서 동일하게 보편적으로 구매하는 카테고리의 가중치를 낮추고, 각 고객 및 군집 별 빈번히 구매하는 상품 품목을 알 수 있다.

월별로 각 고객의 카테고리별 구매 건수를 행렬로 구성한 후, TF-IDF 값을 계산한 결과는 아래의 표와 같다. (전체 행렬 중 일부만 추출하였다.)

고객	스낵류	잎채소	국산과일	탄산음료	냉장조리	일반담배
A	8	5	0	0	0	0
B	0	0	0	1	0	0
C	0	1	0	1	0	0
D	0	0	0	0	1	0
E	0	0	0	0	0	1

[표 2.5-3] TD-IDF 적용 전

고객	스낵류	잎채소	국산과일	탄산음료	냉장조리	일반담배
A	**6.6**	6.59	0	0	0	0
B	0	0	0	2.81	0	0
C	0	2.52	0	2.81	0	0
D	0	0	0	0	3.13	0
E	0	0	0	0	0	**4.02**

[표 2.5-4] TD-IDF 적용 후

두 표를 통해 보편적으로 많이 구매되는 스낵류의 경우 값이 감소하고 특정 고객에게서만 자주 구매되는 일반담배의 경우 값이 커진 것을 확인할 수 있다. 또한 같은 구매건수라도 카테고리별로 적용되는 가중치가 다르다는 것을 확인할 수 있다. 따라서 TF-IDF 적용을 통해 빈도수를 고려한 중요도 값을 얻을 수 있다.

2.5.3 Word Cloud

군집별로 자주 구매되는 카테고리의 특징을 살펴보기 위해 고객별로 구성된 행을 군집별로 묶어 각 카테고리에 대한 평균을 구한다. 표준화와 TF-IDF 과정을 통해 변환된 값을 시각적으로 확인하기 위해 Word Cloud 를 통해 전후의 그림을 비교한다.

<기존 군집 3, 5 의 카테고리별 월별 평균 구매 횟수>

[그림 2.5-1]

<표준화 적용 후 군집 3, 5 의 카테고리별 월별 평균 구매 횟수>

[그림 2.5-2]

<TF-IDF 적용 후 군집 3, 5 의 카테고리별 월별 평균 구매 횟수>

[그림 2.5-3]

먼저 방법론을 적용하기 전에는 두 군집 모두 "스낵류", "국산과일", "잎채소"와 같이 상대적으로 구매횟수가 많은 식품 카테고리가 높은 비중을 차지하고 있음을 확인할 수 있다. 하지만 표준화를 적용한 후 군집 5 번 군에서 "이용권/입장권", "사무기기", "스케이트/썰매"와 같은 새로운 카테고리의 비중이 커진 것을 확인할 수 있으며, TF-IDF 을 적용한 후 군집 5 에서는 "탄산음료", "냉장조리", "생수"의 비중이 커진 것을 확인할 수 있다.

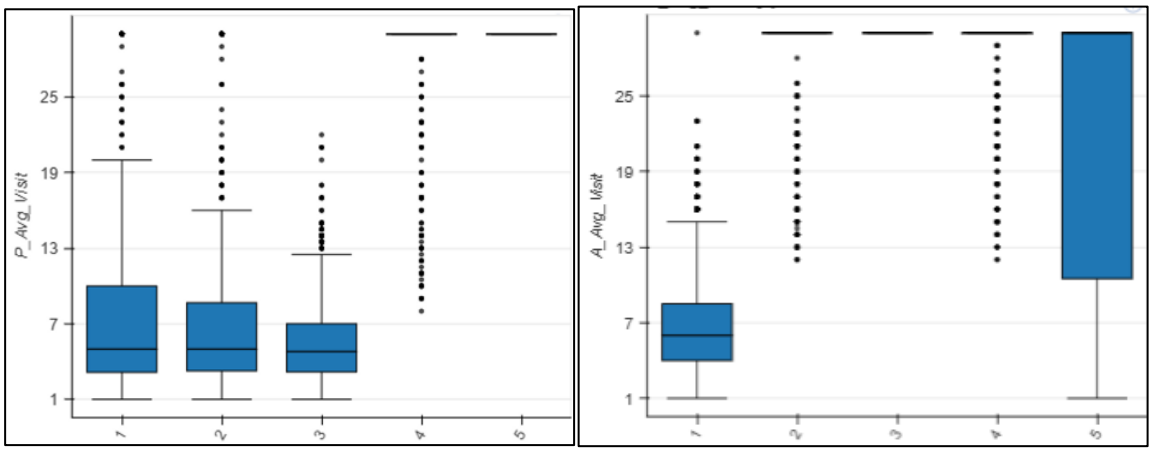
하지만 표준화 방법은 2.5.1 과정을 통해 확인했듯이, 해당 값이 희소할수록 지나치게 높은 값을 부여 받기 때문에 해당 군집에 대한 차별화된 특징으로 볼 수는 있지만, 군집 내 고객들의

보편적인 특징이라고 판단하기 어렵다. 따라서 빈도수를 고려하여 상대적인 중요도를 계산해주는 TF-IDF 방법이 군집 내 고객들의 전체적인 특징을 살펴보기에는 더 적합하다고 판단할 수 있다. 하지만 상대적인 중요도를 위한 가중치가 적용되더라도 보편적인 아이템이 우세할 수 있기 때문에 각 군집의 특징을 파악하기 위해서는 두가지 방법론을 복합적으로 고려해야 한다.

2.6 군집 별 EDA 및 고객 분석

고객이 상품을 구매하고 제휴사를 이용함에 있어서 상이한 행동 패턴을 보인다 가정하였기에, 고객의 상품구매 정보와 제휴사 이용 정보를 각각 eda한 결과를 살펴보았다. 고객의 상품구매 정보와 고객의 제휴사 이용 정보를 통한 변형된 RFM 고객 세분화 지표를 시각화한 결과는 다음과 같다. 군집3의 경우 제휴사를 이용하는 고객이 없으므로 군집3을 제외한 군집의 제휴사 이용 패턴을 분석한다.

2.6.1 평균 구매, 방문 주기 EDA



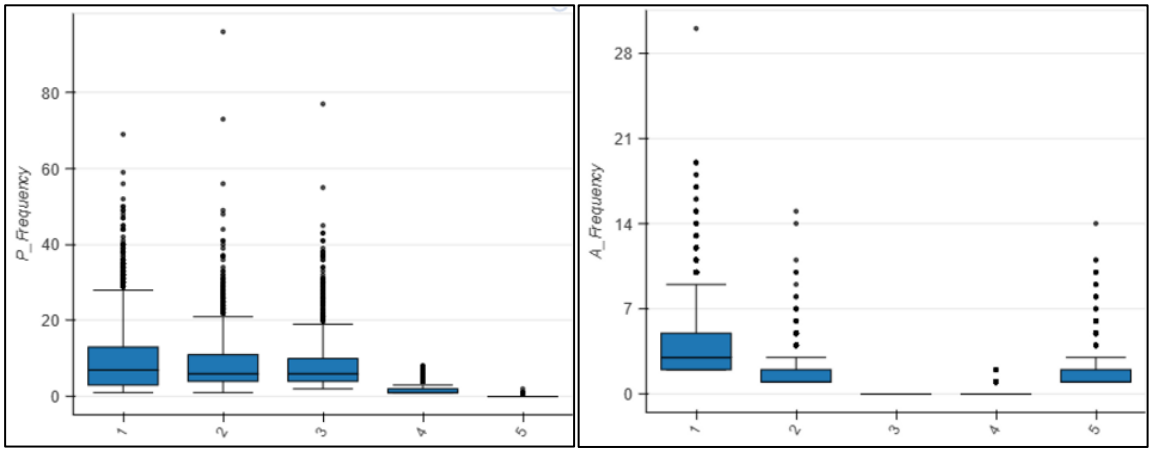
[그림 2.6-1] 왼쪽: 군집 별 평균 상품 구매 주기, 오른쪽: 제휴사 방문 주기

먼저 평균 상품 구매 주기를 나타내는 왼쪽에서 군집4,5의 월별 평균 상품 구매 주기가 평균 26일로 다른 군집에 비해 눈에 띄게 길다는 것을 알 수 있다. 이때 군집5의 경우 극히 적은 상품 구매 이용률로 인하여 평균 상품 구매 주기가 한달을 초과한다. 따라서 군집5의 P_Avg_Visit값을 모두 30으로 설정해주었다. 군집1,2,3은 모두 평균 약 일주일의 짧은 월별 평균 상품구매 주기를 가지고 있으며 표준편차 또한 대략 일주일이다. 그중 군집3의 경우 평균 5일, 표준편차 3일로, 가장 짧은 월별 평균 상품 구매 주기를 보인다. 이는 군집3의 상품 구매가 가장 빈번히 발생함을 의미한다.

다음으로 오른쪽의 제휴사 그림의 경우 군집1의 월별 평균 제휴사 방문 주기가 평균 6일, 표준편차 3일로 다른 군집에 비해 눈에 띄게 짧다는 것을 알 수 있다. 이는 군집3이 가장 자주 제휴사를 방문함을 의미한다. 이때 군집3의 경우 제휴사를 이용하지 않기에 A_Avg_Visit값을 모두

30으로 설정해주었다. 군집2,4,5는 모두 평균 약 22~28일의 긴 월별 평균 제휴사 방문 주기를 가지고 있다.

2.6.2 월별 상품 구매 건수 및 제휴사 이용 건수 EDA

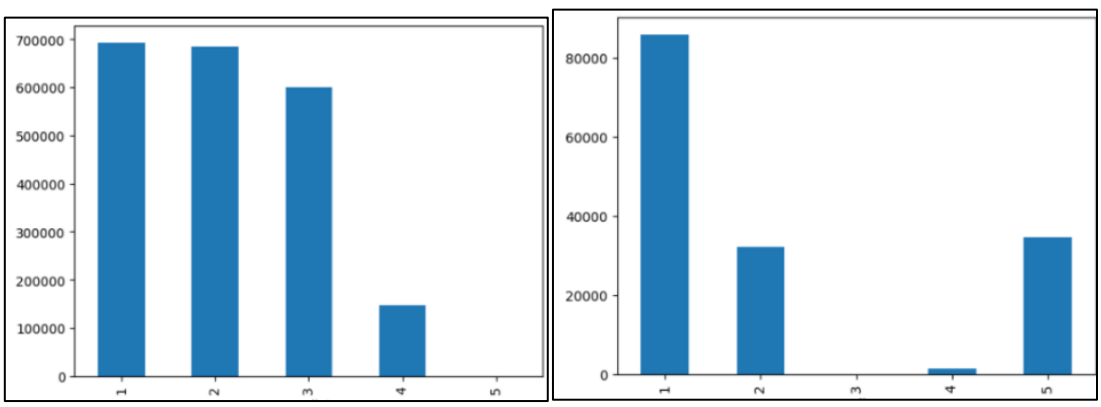


[그림 2.6-2] 왼쪽: 군집 별 상품 구매 건수, 오른쪽: 제휴사 이용 건수

왼쪽 그림에서 군집4,5의 월별 상품 구매 건수는 평균 0,1건로 다른 군집에 비해 눈에 띄게 적다는 것을 알 수 있다. 군집1,2,3은 모두 월 평균 8,9건의 상품을 구매하며 6~9건의 표준편차를 보인다. 그중 군집1의 경우 평균 9건, 표준편차 8건으로 5개의 군집 중 한달 동안 가장 많은 상품을 구매한다.

오른쪽 그림에서 군집4,5의 월별 제휴사 방문 횟수는 평균 4회로 다른 군집에 비해 눈에 띄게 높다는 것을 알 수 있다. 군집2,4,5는 모두 제휴사 서비스를 월 평균 4회 미만 이용한다. 그중 군집4의 경우 월 평균 1회 미만의 극히 적은 제휴사 서비스 이용 횟수를 보인다.

2.6.3 월별 구매 금액 EDA



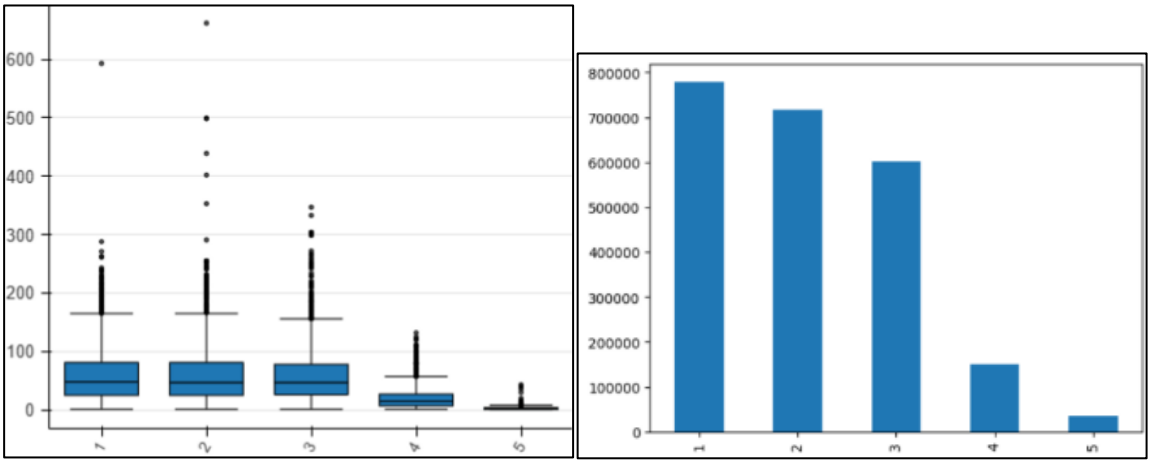
[그림 2.6-3] 왼쪽: 군집 별 총 상품 구매 금액, 오른쪽: 총 제휴사 이용 금액

왼쪽 그림에서 전반적으로 앞서 월별 상품 구매 건수의 eda 결과와 유사한 패턴을 보임을 알 수 있다. 군집4,5의 월별 구매 금액은 평균 15만원 미만으로 다른 군집에 비해 눈에 띄게 적다는 것을 알 수 있다. 군집1,2,3은 모두 월 평균 60~70만원을 소비한다. 그중 군집1의 경우 평균 692,991원, 표준편차 2,616,158원으로 5개의 군집 중 한달 동안 가장 많은 금액을 소비하였다.

오른쪽 그림에서 전반적으로 앞서 월별 제휴사 방문 횟수의 eda 결과와 유사한 패턴을 보임을 알 수 있다. 군집1의 월별 제휴사 이용 금액은 평균 85,917원, 표준편차 620,210원으로 다른 군집에 비해 눈에 띄게 많다는 것을 알 수 있다. 군집2,4,5은 모두 월 평균 3만원을 소비한다. 그중 군집4의 경우 평균 1,441원, 표준편차 6,084원으로 5개의 군집 중 한달 동안 적은 많은 금액을 소비하였다.

2.6.4 상품구매, 제휴사 이용 통합 EDA

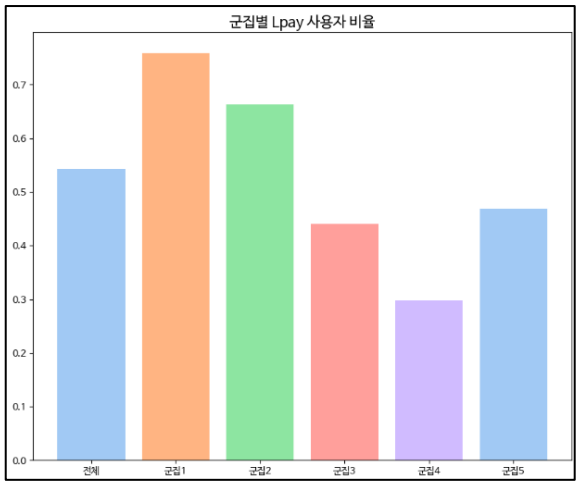
2.6.4.1 월별 구매 수량 & 총 사용 금액

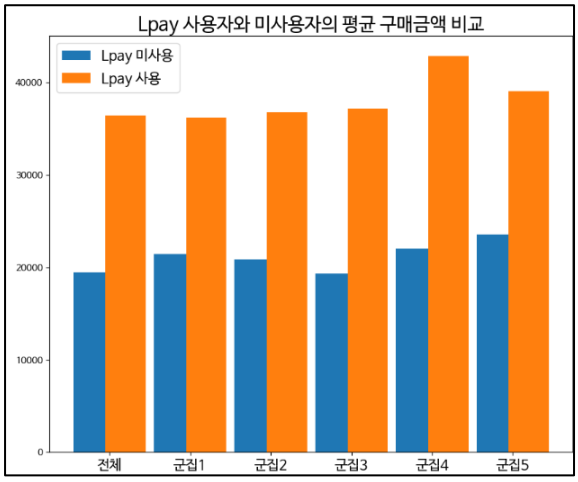


[그림 2.6-4] 왼쪽: 군집 별 총 구매 수량, 오른쪽: 군집 별 총 사용 금액

상품구매에 있어서 전반적으로 군집1,2,3과 군집4,5가 유사한 행동 패턴을 가지고 있다. 월별 구매수량 시각화 결과에서 또한 동일한 패턴을 보임을 확인하였다. 군집1,2,3의 경우 모두 짧은 월별 상품구매 주기와 높은 상품 구매 건수, 금액을 보인다. 군집3의 월별 상품구매 주기가 가장 짧기는 하나, 군집1이 월별 상품구매 건수와 금액에 있어서 더 우수하기에 군집1이 가장 높은 고객 가치를 가진다고 할 수 있다. 제휴사 서비스를 이용함에 있어서도 군집1이 가장 높은 고객 가치를 가짐을 볼 수 있다. 구매 금액이 기업에 있어 가장 중요한 요소라고 판단하여 상품 구매 정보와 제휴사 이용 정보를 통합하여 군집 별 총 이용 금액에 대한 추가적인 eda를 진행하였다. 그 결과, 군집1,2,3,4,5 순으로 가장 많은 금액을 소비하고 있음을 알 수 있다.

2.6.4.2 엘페이 이용률 EDA



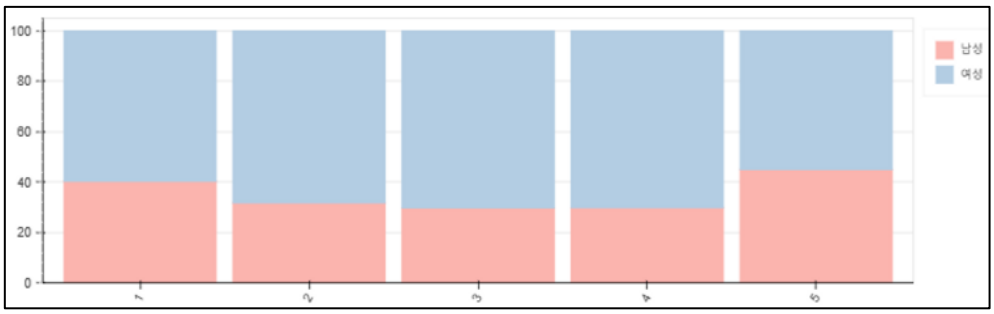


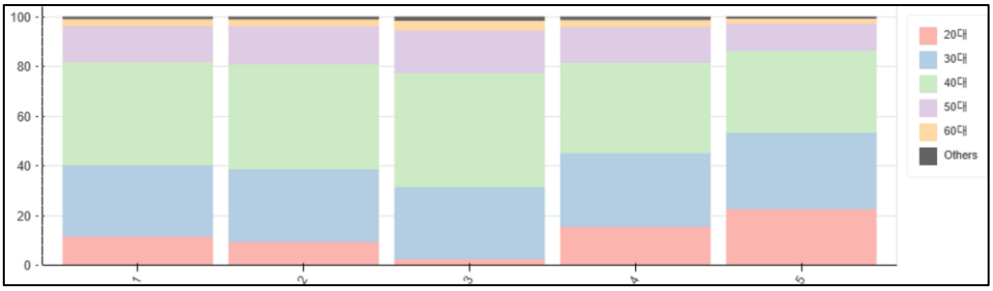
[그림 2.6-5] 왼쪽: 군집 별 엘페이 이용 비율, 오른쪽: 엘페이 평균 구매금액

다음은 군집별 엘페이 이용률을 살펴보았다. 전체와 비교해 보았을 때, 군집1,2의 엘페이 이용률이 높고, 군집3, 4, 5의 이용률이 낮음을 알 수 있다. 엘페이란 롯데멤버스가 제공하는 간편 결제 서비스로, 해당 군집의 엘페이 이용 비율이 높다는 것은 롯데의 이용률이 높다는 것으로 볼 수 있다. 엘페이 사용자와 미사용자의 평균 구매금액을 비교해본 결과, 모든 군집에 있어서 엘페이 사용자의 평균 구매금액이 미사용자보다 높음을 알 수 있다. 따라서 엘페이 사용을 장려하는 마케팅 진행 시 수익 증대의 효과를 얻을 수 있을 것이다.

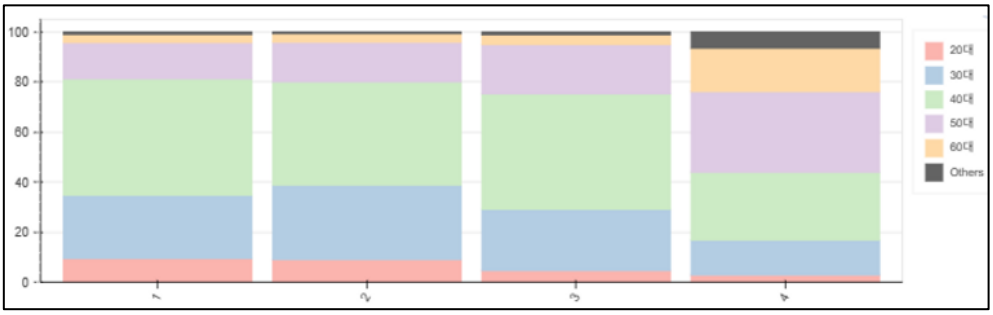
현재 롯데는 자회사의 회원 수 즉, 엘페이 사용자 수를 증진시키기 위한 마케팅 전략을 주로 펼치고 있다. 엘페이 사용자는 구매력이 높은 20~40대가 주를 이루며, 여성 비중이 높은 매체력을 가지고 있다. 엘페이 사용자를 대상으로 미션, 응모 이벤트, 바로적립 등 많은 고객 참여형 서비스를 제공하고 있으며, 이벤트 참여 목적의 캠페인에 있어서 높은 성과를 보인다. 또한 수집한 고객 정보를 바탕으로 보다 적합한 사용자에게 마케팅을 할 수 있는 개인화 마케팅을 가능하도록 한다는 특징을 가지고 있다.

2.6.4.3 추가 EDA 결과





[그림 2.6-5] 제휴사 이용 정보



[그림 2.6-7] 상품 이용 구매 정보

5개의 군집 모두 40,50대의 여성 고객이 주를 이루고 있음을 확인할 수 있다. 상품 구매와 제휴사 서비스 이용 측면에서 모두 군집3은 20대의 분포 비율이 가장 적고, 40대와 60대의 분포 비율이 비교적 높은 편이다. 반면 군집5의 경우 20대가 많은 비율을 차지하고 있음 알 수 있다. 이 외에도 온라인 및 오프라인 여부 정보와 거주지, 이용한 점포 및 제휴사 정보, RFM데이터의 월별 추세 등 다양한 eda를 시도해보았으나 군집에 따른 해당 변수들의 두드러지는 특징은 찾지 못하였다.

2.6.5 고객별 EDA 분석 정리

추가적으로 고객의 상품 구매 정보를 이용하여 거래 품목에 있어서 고객의 구매 패턴을 살펴보았다. 이때 정규화 방법을 통해 구한 군집의 특징과 TF-IDF를 통해 구한 특징을 복합적으로 고려하여 고객의 특성을 파악하였다. 이를 앞서 살펴본 eda결과와 종합하여 군집별로 정리한 결과는 다음과 같다.

1) 군집 1

[그림 2.6-8] 오른쪽: 상품 구매 데이터 속 각 카테고리에 대한 정규화 적용, 왼쪽: TF-IDF 적용

[그림 2.6-9] 제휴사 이용에 대한 TF-IDF 적용 결과

군집 1은 제휴사 이용에 있어서 평균 방문 주기가 다른 군집에 비해 눈에 띄게 짧고, 평균 방문횟수와 구매금액은 월등히 높다는 특징이 있다. 온라인을 이용한 제휴사 이용이 가장 많은 군집이다. 상품구매에 있어서도 낮은 평균 방문 주기를 보이며 평균 구매 건수와 구매금액은 5개의 군집 중 가장 높다. 또한, 엘페이 이용률이 가장 높은 군집으로, '가장 충성 고객'이라고 할 수 있다. 상세 이용 정보를 word cloud를 이용하여 시각화한 결과, 복합 쇼핑몰의 다양한 서비스를 주로 이용한다는 것을 확인할 수 있다. 특히 식당, 카페, 푸드코트를 많이 이용하는 것으로 보아 체류시간이 길 것으로 예상된다.

2) 군집 2

[그림 2.6-10] 왼쪽: 정규화 적용, 오른쪽: TF-IDF 적용

[그림 2.6-11] 제휴사 이용에 대한 TF-IDF 적용 결과

군집 2는 상품 구매에 있어서 군집1과 유사하게 낮은 평균 방문 주기를 가지고 있으며 높은 구매 건수 및 금액을 보인다. 상세 이용 정보를 시각화한 결과에서도 정도의 차이는 존재하나 군집1과 유사한 상품구매 패턴을 가지고 있음을 알 수 있다. 반면 제휴사 서비스 이용에 있어서는 군집1과 상이한 이용패턴을 보인다. 군집1은 모든 제휴사 서비스를 활발히 이용하였지만, 군집2의 경우 복합 쇼핑몰의 카페나 음식점은 군집1과 동일하게 많이 이용하지만 숙박 업종이나 렌탈 업종 등 그 외의 제휴사 서비스는 비교적 적게 이용한다. 다시 말해서 제휴사를 방문하고 소비하기는 하지만 그 규모가 아주 작다. 결론적으로 군집2는 향후 충성고객, 즉 군집1로 전환할 가능성이 가장 높은 고객으로 '충성 후보 고객'이라고 정의할 수 있다.

3) 군집 3

[그림 2.6-12] 왼쪽: 정규화 적용, 오른쪽: TF-IDF 적용

군집 3은 본 논문에서 제시한 5개의 유형 중 가장 속한 고객의 수가 높은 유형으로, 전체 고객의 약 33.1%를 차지하고 있다. 제휴사를 이용하지 않고 오로지 상품 구매만 한다는 특징을 가지고 있다. 40대 여성의 비율이 타 유형에 비해 많으며, 60대의 분포 비율 또한 비교적 높은 편이다. 군집1,2와 유사하게 낮은 평균 구매 주기를 가지고 있으며, 높은 구매 건수 및 금액은 높은 편에 보인다. 군집1,2의 구매 패턴과 다르게 군집 3의 경우 평균 상품 구매 주기는 5개의 군집 중 가장 짧지만 그에 반해 구매 금액은 가장 높지 않다는 특징이 있다. 구매 품목을 살펴본 결과, 청소용품과 계란, 잎채소, 두부/콩나물류, 과일 등 요리와 관련된 품목들을 집중적으로 구매함을 확인할 수 있다. 따라서 식재료와 생활용품을 구매하기 위해 자주 방문하는 '주부'로 구성된 군집이라고 판단하였다.

4) 군집 4

[그림 2.6-13] 왼쪽: 정규화 적용, 오른쪽: TF-IDF 적용

[그림 2.6-14] 제휴사 이용에 대한 TF-IDF 적용 결과

군집 4는 제휴사 이용과 상품 구매에 있어서 평균 방문 주기가 다른 군집에 비해 눈에 띄게 길고, 평균 구매 및 방문 횟수와 금액은 월등히 낮다. 월별 구매 수량 또한 다른 유형에 비해 눈에 띄게 적다는 특징을 가지고 있다. 상품 구매정보를 시각화한 결과, 두드러지는 특징을 없이 다양한 카테고리를 이용하는 것을 확인할 수 있다. 제휴사 이용에 있어서도 다양한 서비스를 이용하는 것을 확인할 수 있다. 따라서 군집4는 필요에 의해서만 해당 서비스를 가끔 이용하는 '충성도가 낮은 고객'이라고 할 수 있다.

5) 군집 5

[그림 2.6-15] 왼쪽: 정규화 적용, 오른쪽: TF-IDF 적용

[그림 2.6-14] 제휴사 이용에 대한 TF-IDF 적용 결과

군집 5는 본 논문에서 제시한 5개의 유형 중 속한 고객의 수가 가장 적은 유형으로, 전체 고객의 약 9.9%를 차지한다. 제휴사를 이용하는 고객이 주를 이루고 있으며, 상품 구매율이 가장 저조하다. 타 유형에 비해 20대의 분포 비율이 높으며, 5개의 유형 중 가장 다양한 제휴사를 비교적 고른 비율로 이용한다는 특징이 있다. 상세 이용 정보를 word cloud를 이용하여 시각화한 결과, 이용권/입장권 등 놀이동산, 여행과 같은 다양한 놀거리와 관련된 키워드들을 발견할 수 있다. 구매 품목에 있어서는 담배, 음료수, 간편식 등 편의점 또는 소형 마트에서 쉽게 찾아볼 수 있는 상품이 주를 이루고 있다. 또한 간편식, 삼각김밥, 라면 등 직접 요리를 해야하는 식재료보다는 조리된 식품을 주로 구매함을 알 수 있다. 따라서 복합쇼핑몰에서 상품을 구매하기보다는 영화 관람 등 '특정 브랜드에 정착하지 않고 문화생활을 활발히 즐기는 젊은 연령대의 고객'이 주를 이루고 있다고 볼 수 있다.

제 3장 고객 유지 및 이탈 고객 관리를 위한 마케팅 전략 방안

앞서 분석한 결과를 통해 이탈 고객과 유지 고객의 인구통계학적, 거래 형태적 특성이 알게 되었고, 이탈에 영향을 끼치는 변수에 대해서 알게 되었다. 또한, 각 군집 별 특징 및 거래 패턴에 대해서도 구체적으로 살펴보았다. 이에 따라 각 군집 별 고객의 이탈 방지 및 재활성화를 위한 고객 관계 관리(CRM: Customer Relationship Management) 프로그램을 전개할 필요성이 있다.

3.1 군집 별 특성에 따른 이탈 방지

EDA를 통하여 클러스터 1의 특성을 알아본 결과 고객들은 다양한 제휴사를 잘 활용하였으며 상품 구매의 빈도수와 금액이 제일 컸음을 알 수 있다. 이를 통해 이들을 해당 기업의 충성 고객으로 정의할 수 있으며 현재의 상태를 유지하고 다른 기업으로 이탈하지 않게 하기 위한 마케팅을 기획해야 할 것이다. 예를 들어서 구매 금액이 큰 고객들이 이용할 수 있는 MVG 혜택을 확대하고 구매 금액과 방문 횟수에 따른 등급을 매겨 롯데에서 구매하는 비율을 더욱더 높이는 프로그램을 실행할 수 있다. 또한, 온라인 쿠폰 및 할인 혜택을 통하여 온라인과 오프라인에서 동시에 마케팅을 진행하여 다양한 방법으로 기업을 이용할 수 있도록 할 수 있다. 클러스터 2의 고객들은 평균 상품 구매 금액이 클러스터 1과 거의 비슷하였으나 제휴사의 이용 빈도에서 차이가 있었다. "제휴사를 많이 이용할수록 이탈률이 떨어진다"는 분석 결과에 따라서 이들에게는 기존에 주로 이용하고 있던 음식 관련 제휴사가 아닌 다양한 제휴사를 이용할 수 있는 기회를 주는 프로그램을 기획할 필요성이 있다. 예를 들어 현재 롯데에서 진행하고 있는 엘스탬프, 롯키데이와 같은 제휴사 통합 서비스를 집중적으로 적용하여 군집 2 고객들의 새로운 제휴사 사용을 증가시키고 롯데에서의 다양한 경험을 늘려갈 필요성이 있다. 클러스터 3의 분석 결과 총 고객 수가 가장 많았으며 평균 방문 기간이 제일 짧았지만 총 구매 금액은 평균임을 알 수 있었다. 또한, 타 그룹에 비하여 40대 여성의 비율이 압도적으로 많았으며 식품 코너 이용 비율이 컸음을 알 수 있다. 이를 통해 기업은 클러스터 3을 주부로 초점을 맞추고 마케팅을 해야 한다. 식품 및 의식주 맞춤 마케팅이나 쿠폰 할인 이벤트를 통해 40대 여성들이 롯데의 식품관을 더 많이 이용할 수 있도록 하여야 한다. 또한, 문화 센터 및 다양한 프로그램을 이용하여 타 기업을 이용하는 유사 고객들의 유입을 증가시키고, 이들이 롯데의 공간에 오래 머무를 수 있도록 하여 이탈을 방지해야 한다. 또한, 이들은 제휴사를 아예 방문하지 않는 군집으로, 상대적으로 진입 장벽이 낮은 음식점이나 엔터테인먼트01 제휴사를 이용하도록 유도하여 롯데의 다양한 서비스를 경험할 수 있도록 유도하는 마케팅이 필요하다.

클러스터 4를 분석한 결과 평균 방문 기간이 가장 길고 평균 방문 빈도수와 구매 금액이 다른 유형에 비해 눈에 띄게 낮음을 알 수 있다. 해당 클러스터 고객은 특별한 특징이 없으며 필요할 때만 롯데 서비스를 이용하는 충성도가 낮은 고객임을 알 수 있다. 따라서 이들에게는 첫 구매 이벤트나 쿠폰 마케팅을 이용하여 롯데를 이용할 시 즉시 혜택을 받을 수 있도록 하여 구매 촉진을 일으킬 수 있을 것이다. 마지막으로 클러스터 5의 결과 20대가 차지하는 비율이 가장 높으며, 1번 군집과 비슷하게 혹은 오히려 더 다양한 제휴사 서비스를 이용하고 있다. 이들은 이미 다양한 유통사를 이용하고 있으며 기업에서 제공하는 이벤트를 잘 활용하고 있으므로 개개인 맞춤 홍

보 마케팅을 통하여 이용을 촉진할 수 있다. 또한, Y 커뮤니티와 같은 경험을 통하여 기업에 대한 인식을 좋게 만드는 것과 함께 상품 구매를 유도하여 미래의 충성 고객이 될 수 있게 만드는 전략이 필요하다.

3.2 이탈률이 작은 군집으로 이동을 위한 고객 관리

경쟁 사회가 심화되고 있는 현재, 기업들은 새로운 고객을 획득하는 것보다 기존의 고객들이 이탈을 하지 않고 유지하는데 더 많은 노력을 들이고 있다. 비용적인 측면에서도 기존 고객 유지가 신규 고객 확보보다 이득이라는 것은 상식이다. 따라서 본 연구에서는 이탈률이 큰 4, 5번 군집이 1,2,3번 군집으로 이동하게 하는 것이 중요한 포인트이다.

이탈률이 큰 군집에서 작은 군집으로 사람들을 움직이게 하기 위해서는 재방문에 중점을 두어 마케팅을 해야 한다. 고객이 재방문을 한다는 것은 고객이 원하는 물건이나 서비스가 롯데에 존재해야 하고, 고객이 원하는 상품이 롯데에만 존재한다면 재방문의 확률이 더욱더 높아질 것이다. 롯데 APP이나 Lpay를 통하여 고객이 구매하고자 하는 상품이 롯데에 있음을 알려주는 알림 서비스를 제공한다면 고객의 재방문을 유발할 수 있을 것이다. 또한, "보틀 벙커"와 "프리미엄 카레"와 같은 롯데에서만 구매할 수 있는 상품을 구비하여 한번 이용한 고객이 다른 기업으로 이탈하지 않고 재방문을 하게하는 마케팅이 필요하다.

기존의 일반 고객들을 충성 고객으로 만드는 것은 새로운 고객을 유치하는 것과 달리 긴 시간이 필요하며, 점진적인 이익을 기대하여야 한다. 따라서 이탈률이 높은 등급을 한 번에 상위 그룹으로 이동하기보다 단계적인 마케팅을 통한 성장이 중요하다. 롯키데이와 스노우포인트와 같은 서비스를 통하여 첫 구매 고객의 재방문을 도모하고, Lpay 프리미엄과 같은 마케팅으로 롯데 제휴사의 이용 빈도를 늘리고, 기존 구매자에게 업그레이드된 상품을 판매하는 "상품 업셀링"으로 고객의 구매 규모를 늘리는 것과 같은 점진적인 마케팅을 진행하는 것이 중요하다. 위와 같은 마케팅을 통하여 이탈률이 높은 고객이 롯데의 충성 고객으로 전환된다면 기업에게는 경제적으로 큰 이득이 있을 것으로 예상된다.

제 4장 결론

과격화 된 경쟁 속 신규 고객의 유입뿐만 아니라 기존 고객의 성공적인 관리가 이루어져야 한다는 성공적인 고객 관리(CRM)의 필요성이 대두되고 있다. 따라서 본 연구는 시변성을 가지고 있는 고객의 구매 이력을 이용하여 고객을 세분화하고 이탈률에 따라 등급을 부여하는 새로운 방법을 제안함으로써 효과적인 고객 가치 분석에 도움을 주고자 한다. 추가적으로 군집 별 고객의 이탈 예측 모델을 제시함으로써, 이를 바탕으로 개인의 행동 패턴을 고려한 고객 별 차별화된 마케팅 전략을 제공하기를 제안한다.

본 연구에서는 12개월 동안의 고객의 구매이력 데이터를 이용하였으며, 고객 세분화, 이탈 예측, 고객 분석 3단계로 나누어 전체적인 모델을 구성하였다. 본 논문에서 제안하는 새로운 RFM 지표를 기준으로 고객을 K-Means Clustering을 통해 5개의 군집으로 세분화한 후, 군집별 이탈률과 월별 이탈률의 변동을 확인하여 1에서 5까지 군집에 등급을 부여하였다. 고객 특성을 비롯하여 새로운 RFM지표의 시계열적인 특성을 반영하기 위하여 이를 2,3,4,5,6개월의 단위로 그 기준을 변화시키며 이탈 예측 모델의 성능을 평가하였다. 그 결과 6개월이 가장 최적의 반영 월 수로 판단되었으며, 이에 따른 최종 모델은 test set에 대해 81.04%의 정확도를 가진다. EDA를 비롯하여, 표준화를 통한 카테고리별 비교와 TF-IDF방법을 활용한 상대적인 중요도로 군집 별 특징적인 아이템을 파악함으로써 군집 별 고객의 특징을 도출하고 이탈 등급을 고려한 맞춤화 마케팅 전략을 제안하였다.

본 논문은 다음과 같은 의의를 갖는다. 첫째, 기존의 RFM(Recency, Frequency, Monetary)가 아닌 제휴사 이용과 상품 구매 데이터를 분리하여 각각의 평균 이용 주기, 건수, 금액을 살펴보는 변형된 새로운 RFM을 제안하였다. 이는 고객이 제휴사를 이용하고 상품을 구매하는 행동 패턴에 있어서 상이한 행동 패턴을 보일 수 있다는 가정 하에 고객의 상품 구매와 제휴사 이용을 분리하여 살펴봄으로써 각각의 행동 패턴을 파악하였다는 점에서 의의가 있다. 또한 기존의 Recency 변수를 평균구매주기 변수로 대체함으로써 빠르게 변하는 고객의 행동 패턴을 월별로 파악한 후 이를 반영하였다는 점에서 기존 논문과 차별화된다. 둘째, 이탈 가능성이 높은 군집과 충성고객의 구매 특성을 세분화하여 군집별 이탈 등급 파악하였다. 셋째, 월별 Avg_VFM과 3가지 고객 특성 변수 데이터만으로 이탈예측 모델을 구축함으로 모델의 간결성 확인하였다. 또한 비교군을 통해 Avg_VFM 값의 시계열적인 변화를 고려한 모델의 이탈 정확도가 더 높다는 것을 확인하였다. 넷째, 군집 별 구매 품목에 따른 특징 파악에 있어서 TF-IDF와 Standardization을 사용함으로써 각 군집 속 일부 고객의 차별화된 아이템과 각 군집을 대표적인 특징을 나타내는 아이템을 도출하였다. 따라서 해당 결과와 군집 별 EDA를 통해 고객 세분화를 통한 고객 유지 및 이탈 방지 마케팅 방안을 제안하였다.

분석을 통해 도출한 5개의 군집의 특징과 그에 따라 제안하는 마케팅 방안은 다음과 같다. 군집1은 제휴사와 상품구매를 모두 활발히 이용하며 소비 금액 또한 가장 많은 '충성 고객'이다. 따라서 MVG 혜택을 확대하고 구매 금액에 따라 등급 부여하는 등의 마케팅 전략을 통해 이탈을 방지하고 현 등급을 유지하는 것이 중요하다. 군집2는 충성고객인 군집1과 유사한 행동 패턴을 보이나 제휴사를 잘 활용하지 못하는 '충성 후보 고객'이다. 롯키데이. 엘스탬프와 같은 제휴사 통합 서비스를 집중적으로 활용하여 다양한 제휴사를 이용할 수 있는 기회를 마련해주는 것을

제안한다. 군집3은 본 논문에서 제시한 5개의 유형 중 가장 속한 고객의 수가 높은 유형으로 제휴사 이용없이 상품만 구매하며 '주부'가 주를 이루고 있다. 식재료와 생활용품을 빈번히 구매한다는 점을 고려하여 식품관 할인 이벤트 알림 서비스 및 문화센터 등의 공간 활성화를 통해 체류시간을 늘리는 방안을 추천한다. 군집4는 '충성도가 낮은 고객'으로 이용률이 저조하며 소비 금액 또한 월등히 낮다. 따라서 첫 구매 시 혜택을 주는 마케팅을 이용함으로써 전체적인 이용을 촉진시키고 재방문을 도모하는 마케팅 방안이 필요하다. 군집5는 다양한 문화생활을 제공하는 제휴사의 이용은 활발한 편이나 그에 반해 상품 구매 이용률은 저조하다. 5개의 유형 중 젊은 세대가 가장 많은 비중을 차지하는 만큼, 개개인 맞춤 마케팅 활성화 및 다양한 경험을 통하여 미래의 충성 고객으로 유도하는 방안을 제안한다.

본 논문에서 제시한 모델을 통해 높은 이탈률을 가진 고객이 충성 고객으로 전환할 수 있도록 맞춤형 마케팅 전략을 취한다면, 기업은 큰 경제적 이익을 얻을 수 있을 것으로 기대된다.

참고문헌

(2002)/"CRM핵심모듈 고객이탈방지 툴 기업전략 부재 효과 못 거둬," 디지털타임즈, 2002. 08. 07/김응렬

(2002)/"신용카드 시장에서 데이터마이닝을 이용한 이탈고객 분석", 성균관 대학교/이건창, 정남호, 신경식

(2010)/ "Predicting customer churn in mobile networks through analysis of social groups."/ Richter, Yossi, Elad Yom-Tov, and Noam Slonim.

(2022)/"기계 학습과 RFM 모형을 이용한 고객 유형 분석 – 문화예술 이용객을 대상으로"/중앙대학교 대학원 경영학과 경영과학 생산관리 경영정보시스템 전공 박지현

(2005년 4월)/"CRM 고객데이터 분석을 통한 이탈고객 연구" / 김상용, 송지연, 이기순

(2003)/"데이터마이닝을 활용한 동적인 고객분석에 따른 고객관계관리 기법"/하성호, 이재신

(2017)/"Predicting Customer Churn Using Recurrent Neural Networks"/ JESPER LJUNGEHED

(2006년 5월)/"균형적 고객세분화에 관한 사례 연구"/윤종육, 윤종수

(2010)/"의사결정나무 기법을 활용한 백화점의 고객세분화 사례연구"/채경희, 김상철

(2014)/"항공사 고객가치 기반의 고객세분화 사례연구"/박광식

(2013)/"RFM에서 등급부여 방법에 관한 연구. 한국데이터 정보과학회지"/류귀열, 문영수

(1994)/"Strategic database marketing. Chicago: Probus Publishing Company"/Hughes

(2007)/"고객의 행동 변화를 통한 신규고객 세분화와 구매항목 예측"/도희정, 김재련

(2001)/"은행고객 세분화를 통한 이탈고객 관리 분석- 가계성 예금을 중심으로"/이건창, 권순재, 신경식

(2012)/"고객지향 세분시장 획득을 위한 데이터 마이닝 기법 적용방안"/김종호

(2017) Mohammadzadeh, Mehdi, Zeinab Zare Hoseini, and Hamid Derafshi. "A data mining approach for modeling churn behavior via RFM model in specialized clinics Case study: A public sector hospital in Tehran." Procedia computer science 120

(2023) ZELENKOV, Yury A.; SUCHKOVA, Angelina S. Predicting customer churn based on changes in their behavior patterns. Бизнес-информатика,

(2013) 유귀열; 문영수. RFM 에서 등급부여 방법에 관한 연구. 한국데이터정보과학회지, 2013, 24.2: 245-255.

MODEO, R. F. M. RFM 모델 기 반의 병원고객 세분화 전략.

(2010) 채경희; 김상철. 의사결정나무 기법을 활용한 백화점의 고객세분화 사례연구. 유통과학연구

(2020) 나광택, et al. 증권 금융 상품 거래 고객의 이탈 예측 및 원인 추론. 한국빅데이터논문지 제

(2020) 김형수; 홍승우. 이차원 고객충성도 세그먼트 기반의 고객이탈예측 방법론. 지능정보연구

(2018) 김경태; 이지형. 딥 러닝과 Boosted Decision Tree 를 활용한 고객 이탈 예측 모델. 한국지능시스템학회 논문지

(2016) 이세희; 이지형. RNN 을 이용한 고객 이탈 예측 및 분석. 한국컴퓨터정보학회 학술발표논문

부록

[부록 1] Hyper -parameter 값

개월 수 길이 비교에 대한 통일된 하이퍼파라미터 값

모델	Hyper-parameter
Logistic	C= 0.08786879644302269, penalty ='l2'
XGB	subsample= 0.9, n_estimators= 200, max_depth= 7, learning_rate= 0.1, gamma= 0, colsample_bytree= 0.7
RF	bootstrap= False, max_depth= 19, max_features='sqrt', min_samples_leaf= 3, min_samples_split=6, n_estimators= 918
CatBoost	border_count=64, depth= 8, iterations=786, l2_leaf_reg=3, learning_rate=0.5
GBM	learning_rate= 0.1, max_depth= 7, n_estimators=300
K-NN	n_neighbors= 15, p= 1, weights= 'distance

[월 비교 모델 별 학습 파라미터]

[부록 2] GridSearchCV hyper-parameter Tune-grid

RandomForestClassifier 에 대한 GridSearchCV 에 사용한 값은 다음과 같다.

GridSearchCV Tune-Grid	
Hyper-parameter	**Grid**
n_estimators	**[100,150,300,500,1000]**
Min_samples_split	**[2,3,5,10]**
Min_samples_leaf	**[1,2,3,5,10]**
Max_depth	**[5,15,30,50]**
Max_feature	**['sqrt', 'log2']**
BestTune:{ "n_estimators"=100, "Min_samples_split"=2, "Min_samples_leaf"=1, "Max_depth"=50	

[GridSearchCV Tune-Grid]

Max_feature 의 경우 모델을 학습하고 Test 를 예측하는 과정에서 'log2'의 결과가 더 좋았다. 또한 그 외 변수들은 각 Default 값에서 가장 결과가 좋았으며, 최종 적으로 위 값에서 Max_depth 까지 삭제한 기본 모델의 Cross-Validation 의 정확도가 가장 높게 특정 되었다.

[부록 3] Recency 변수 추가 전후 Vif 값 변화

변수 명	Vif	Recency 추가 후	Vif
6Pre_A_Visit	2.73	6Pre_A_Visit	10.14
5Pre_A_Visit	2.73	5Pre_A_Visit	9.62
4Pre_A_Visit	2.74	4Pre_A_Visit	9.31
3Pre_A_Visit	2.70	3Pre_A_Visit	9.14
2Pre_A_Visit	2.66	2Pre_A_Visit	8.93
Pre_A_Visit	2.65	Pre_A_Visit	8.89
		6Pre_A_Recency	6.42
		5Pre_A_Recency	5.99
		4Pre_A_Recency	5.84
		3Pre_A_Recency	5.78
		2Pre_A_Recency	5.59
		1Pre_A_Recency	5.67