소비패턴을 통한 보험종목별 사고율 분석

목차

○ 개요

1. 요약

- 카드 매출 데이터를 군집화해서 카드 소비 집단별 특성을 파악해 소비 패턴을 파악한다. 파악한 소비 패턴을 이용해 보험 종목별 사고율 분석을 시행한다. 이때, 사고율 분석을 위해 GLM(일반화선형모형)을 도입한다. 분석을 통해 얻은 인사이트를 통해 삼성카드와 보험개발원에서 활용할 수 있는 아이디어를 제시한다.

2. 착안점

1) 카드 소비 데이터 분석을 통해 군집별 소비 특성을 도출한다.

2) 군집별 소비 특성을 토대로 보험 종류별 사고율을 분석한다.

3. 문제 해결 과정

1) 연구 계획단계에서는 사고율 데이터로 예측 모델링을 진행하려고 했으나, 자동차보험 사고율이 모두 결측치인 관계로 우선 해당 결측치에 대한 처리가 필요했다. 이를 위해 장기보험 사고율 데이터로 학습시킨 머신러닝 모델을 사용해 자동차보험 사고율을 예측했다.

2) Poc 제작 시 대용량 데이터일 것으로 예상해 MZ세대의 데이터만을 활용한 분석을 하려고 했으나, 본선 데이터셋은 데이터 사이즈가 크지 않아 전체 데이터를 활용했다. 대신 '**MZ세대 여부**'라는 파생변수를 만들어 해당 세대가 많이 속한 클러스터의 특성을 파악했다.

3) 생명보험의 '**사고건수**' 변수에 값이 0인 데이터가 지나치게 많아(16,377건 중 16,360건, 약 99.9%) 사고율 추정 모델의 계산이 어렵다는 문제가 발생했다. 이에 감마분포를 바탕으로 모델링을 진행하고자 했으나, 분포 추정의 문제가 발생했다. 이를 해결하기 위해 사고율의 계산식을 이루는 두 변수 '경과계약건수'와 '사고건수'가 이산형 자료임에 착안했다. 각 변수에 대한 일반화 선형 모형(Generalized Linear Model, 이하 GLM)을 각각 추정한 후, 이를 사고율의 계산식에 대입해 최종적으로 사고율을 설명하는 회귀식을 추정하는 방향으로 이를 해결하고자 했다.

$$\widehat{f_{\text{사고율}}} = \frac{\widehat{f_{\text{사고건수}}}}{\widehat{f_{\text{경과계약건수}}}}$$

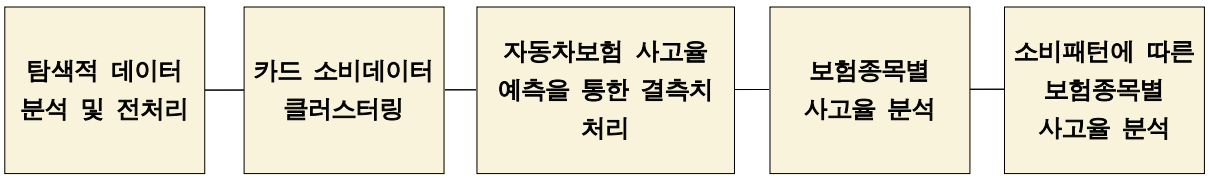
4) Poc 제작시 머신러닝과 딥러닝을 활용해 보험 종목별 사고율 예측 모델링을 진행하려고 했다. 하지만 본선 데이터셋 확인 후, 데이터셋의 사이즈가 작아 학습량이 적어 머신러닝과 딥러닝 모델 성능이 낮을 것으로 판단했다. 따라서 데이터 증폭을 고려해보았다. 사고율과 같은 연속형 변수를 증폭할 시 MC(Monte Carlo) 샘플링이나 Bootstrap 샘플링을 기반으로 한 방법을 고려할 수 있다. 우선 MC 샘플링의 경우 사고율에 대한 분포가정이 필요하다. 사고율이 감마분포를 따른다는 것은 선행연구를 통해 파악했으나, 감마분포 모수 추정 시 log-likelihood function이 극한에서 잘 수렴하려면 데이터 수가 충분해야 했다. 그러나 사고율이 0이 아닌 데이터가 매우 적어 추론 과정에서의 MLE(Maximum Likelihood Estimator) 값의 consistency가 위배되어 미흡한 추정량이라고 판단해 MC 샘플링은 사용하지 않았다. 분포가정 없이 복원추출을 기반으로 데이터를 증폭하는 bootstrap 기법을 고민해보았다. 하지만

복원추출 대상인 사고율 데이터가 적고 매우 편향되어 있어 해당 기법을 적용할 경우 입증된 사고율 분포인 감마분포 등과 멀어져 현실설명력이 떨어질 수 있다고 판단했다. 따라서 데이터셋을 증폭하지 않고 원 데이터 그대로 통계적 모델링을 진행했다.

○ 주요기술

1. 차원축소 후 클러스터링: 차원 축소 기법 중 하나인 PCA(주성분분석, Principal Component Analysis)는 분산을 최대한 유지하며 축의 개수를 선택할 수 있기 때문에 군집분석 시 정보량이 보존되어 적절할 것이라고 판단해 사용했다. PCA 적용 후, Kmeans 클러스터링 기법을 활용해 군집분석을 진행했다.
2. LightGBM: 트리기반 머신러닝 알고리즘으로, leaf-wise 분기방식을 사용한다. LightGBM을 선택한 이유는 학습시간이 빠른 특징이 있어 제한된 분석환경 내에서 신속하게 최적화하고 모형을 적합시킬 수 있는 트리모델이 필요했기 때문이다.
3. GLM(일반화선형모형, Generalized Linear Model, 이하 GLM): 선형회귀모형을 확장시킨 모형으로, 종속변수가 정규분포를 따르지 않는 경우에 사용한다. GLM은 종속변수의 기댓값을 연결함수로 변환한 독립변수들의 선형결합으로 표현할 수 있다. GLM을 사용한 이유는 종속변수가 정규분포를 따르지 않는 상황에서 변수의 영향력을 잘 설명할 수 있는 모형이기 때문이다.

○ 데이터 분석 및 모델링 과정



1. 전처리
2. 카드 소비데이터 클러스터링
3. 자동차 사고율 예측을 통한 결측치 처리
4. 보험종목별 사고율 분석
5. 소비패턴에 따른 보험종목별 사고율 분석

○ 데이터 분석 및 모델링 방법론

1. 전처리

- 본격적인 데이터 전처리에 앞서, 초기 데이터가 총 85,001개의 행과 118개의 변수로 구성되어 있음을 확인했다. 또한 분석 상 편의를 위해, 모든 영어 변수명은 한글로 변환하여 진행했음을 밝힌다.

(1) 결측치 처리

- 열별 결측치를 확인한 결과, 보험료 및 손실액에 상당 부분 결측치가 존재하며, 해당 부분은 모두 자동차보험(보험종목 = 3)의 보험료 및 손실액인 것으로 파악했다. 변수별 결측치의 수는 다음과 같다.

변수명	결측치 수
손해액	52970
자동차국외산구분코드	30791
자동차보험가입경력	30791
보험료	54210
거주지_시군구	697
직장지_시군구	442

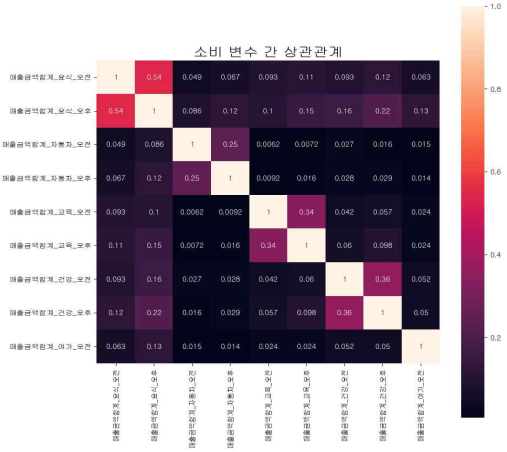
- 이때, 손해액 및 보험료의 경우 분석 주제인 '사고율'과 직결되는 항목이기에, 타 보험의 항목을 참고해 적절히 대체했다.
- '거주지_시군구' 및 '직장지_시군구' 변수의 경우, 지나치게 지엽적인 정보라고 판단해 해당 변수를 제거했다.
- '자동차국외산구분코드' 및 '자동차보험가입경력' 변수의 경우, 보험 종목의 특성상 자동차보험 이외의 항목에서는 결측치가 많이 존재할 수밖에 없으므로 결측치 대체를 진행하지 않고 이외의 보험 종목에 대해서만 해당 변수를 제거했다.

(2) 변수 삭제 및 파생변수 생성

- 기준년도 변수의 경우, 모든 데이터에 대해 동일한 값(2021)만 존재하기에 분석에 영향을 미치지 못할 것으로 보고 삭제했다.
- 각 보험 종류에 따른 사고율 계산식을 통해 '사고율' 변수를 생성했다.
- 소비자 연령대에 따른 세대 구분이 사용자의 소비패턴 및 보험사고율 여부에 중요 영향을 미칠 수 있을 것으로 파악해, 45세를 기준으로 'MZ여부' 변수를 추가했다. (45세 이하 = 1, 45세 초과 = 0)
- 소비자의 품목별 소비패턴 변수에 대해, 소비패턴을 명확히 파악하고자 다음과 같은 파생변수를 생성했다.

  · 데이터에 존재하는 시간적인 특성을 제거하고 데이터의 정보를 요약해 제시하기 위해 아래의 두 가지 파생변수를 생성해 기존의 매출 특성 데이터를 대체했다.

  · 각 매출기록에 대한 오전, 오후 변수가 상호 간에 대해 선형 관련성이 높을 것으로 예상했으나, 상관관계 분석 결과 유의미한 관련성이 나타나지 않았다.



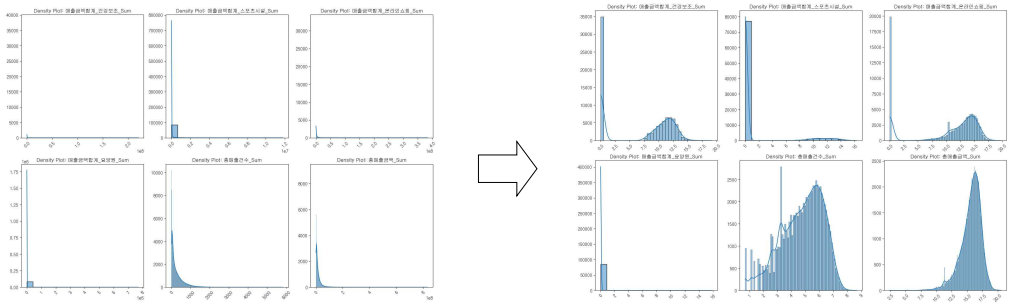
파생변수명	변수 설명
매출건수_Sum	각 품목의 오전과 오후의 매출건수를 더한 변수
매출금액_Sum	각 품목의 오전과 오후의 매출금액을 더한 변수
매출건수_오전여부	매출 건수가 오전과 오후 중 어느 시간대에 많은지를 나타내기 위한 변수
매출금액_오전여부	매출 금액이 오전과 오후 중 어느 시간대에 많은지를 나타내기 위한 변수
총_매출건수_Sum	전체 매출건수의 총합
총_매출금액_Sum	전체 매출금액의 총합

(3) 범주형 변수 처리 (Encoding)

- 이상의 전처리 과정을 거친 후, 다음 5개의 변수가 문자열 자료형임을 확인했다.

['추정_직업군', '추정_연소득', '거주지_광역', '직장지_광역', '추정_라이프스테이지']
- 해당 문자열 자료들의 고유값 수가 적고(20개 미만), 포함하는 정보가 범주형 자료임을 파악해 변수별 적절한 인코딩을 진행했다.
- '거주지_광역', '직장지_광역' 변수의 경우, 기존의 자료는 포함하고 있는 정보가 지엽적이라고 파악해 수도권 여부를 기준으로 거주지 및 직장지의 소재지가 수도권인지에 따라 binary 인코딩을 진행했다.
- 이외의 모든 변수는 라벨 인코딩을 통해 숫자로 표현된 범주형 변수로 변환했다.

(4) 데이터 변환(Scaling)
- 초기 데이터 분포를 파악한 결과, 대다수의 연속형 변수가 왼쪽으로 치우친 분포를 보여, 대부분의 변수에 이상치가 존재함을 파악했다.
- 수치형 데이터에 로그 변환을 취한 후의 그래프를 살펴본 결과, 분포의 쏠림 현상을 완화할 수 있었다.

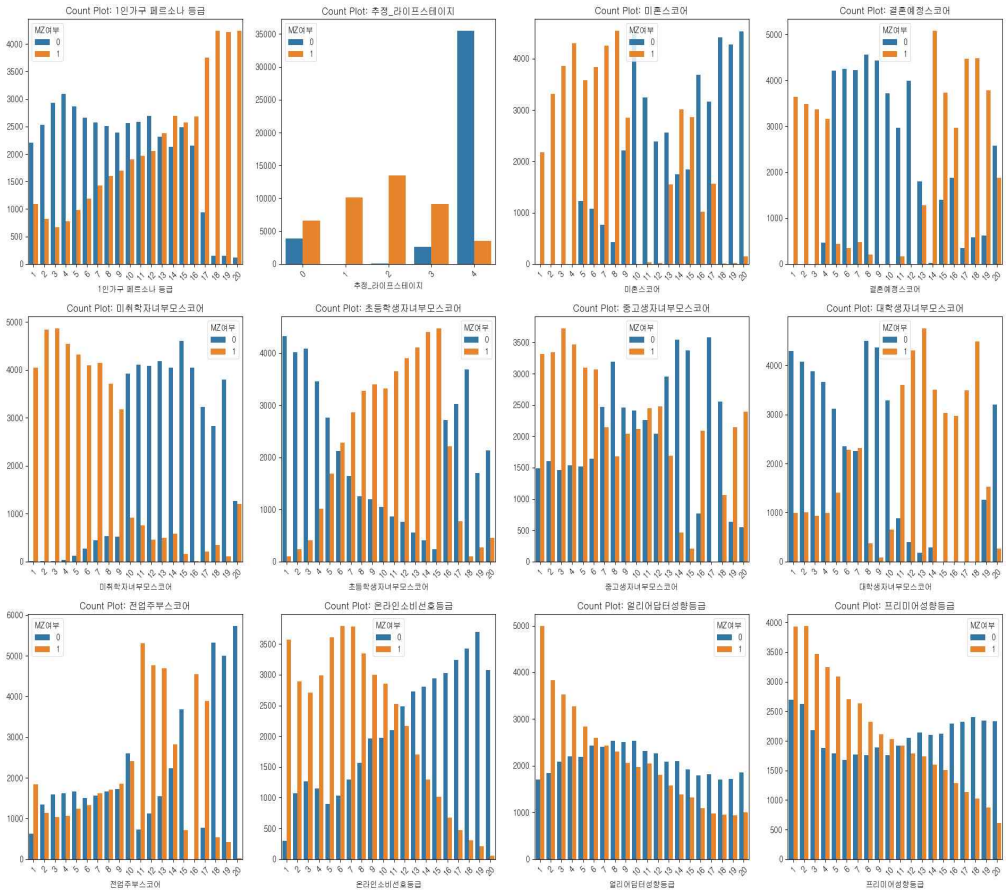


- 로그 변환 후 추가적으로 표준화(Standard Scaling)을 진행했다. 일부 변수에 대해서는 표준화가 데이터의 변수 형태를 유의미하게 변형시켰으나, 이외의 변수에 대해서는 유의미한 성과를 보이지 못했음을 확인했다. 이는 해당 변수에 대해 동일값이 지나치게 많고, 극단적인 이상치 값이 존재해 변수의 범위를 줄이는 로그 변환이 제대로 작동하지 못했기 때문으로 추정된다. 따라서 표준화는 군집분석의 차원축소 과정에서만 사용했다.

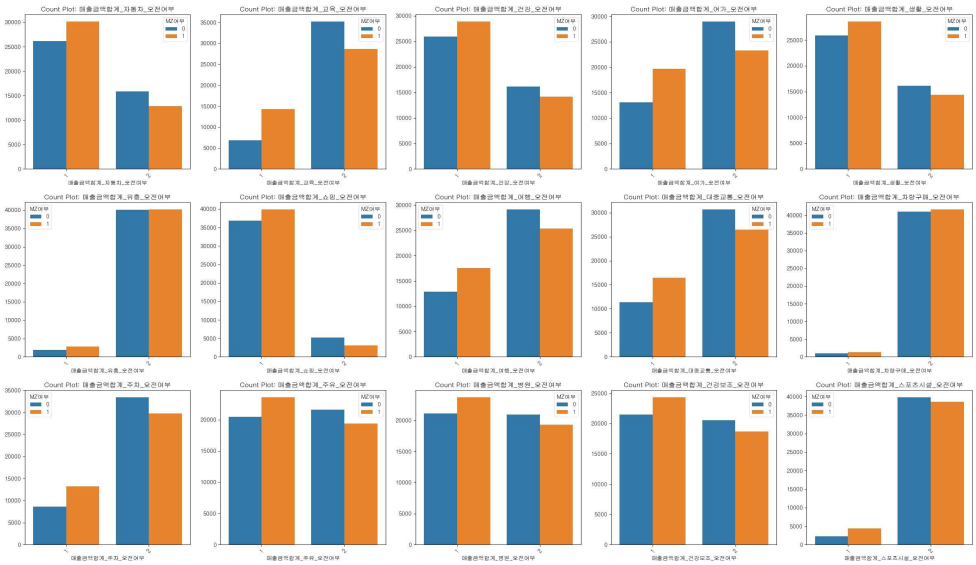
2. EDA

- 앞의 과정을 통해 생성된 데이터를 바탕으로 소비자의 일관된 매출 특성을 파악하고자 추가적인 분석을 진행했다.

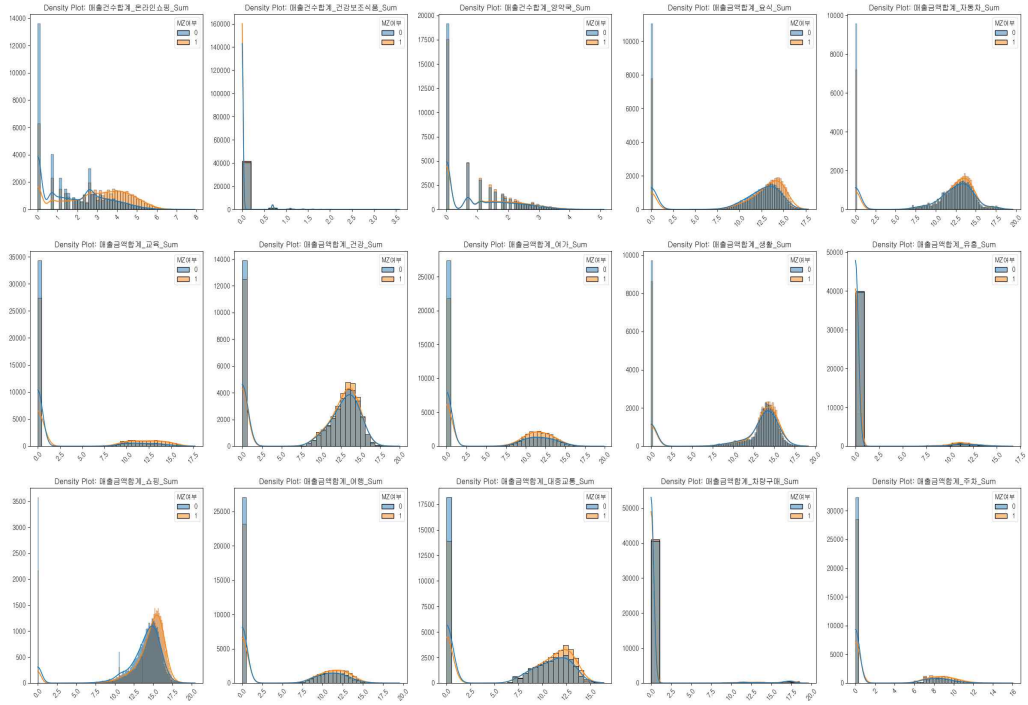
(1) 연령에 따른 분석
- 다음은 'MZ여부' 파생변수에 따른 각 변수별 분포를 시각화한 것이다.
- 인구통계학적 특성을 나타내는 스코어 관련 데이터의 경우, MZ세대와 기성세대는 대부분의 영역에서 상반된 분포를 보였다.



- 또한, MZ세대는 전 라이프스테이지에 고르게 분포한 반면, 기성세대는 대부분의 인구가 중고생 자녀를 두고 있는 것으로 드러났다.
- 이를 통해 MZ세대와 기성세대의 전반적인 관심사 및 생활패턴이 상반된다는 것을 파악할 수 있었다.
- MZ세대는 기성세대에 비해 오전에 더 많은 소비를 하는 경향이 있었다.

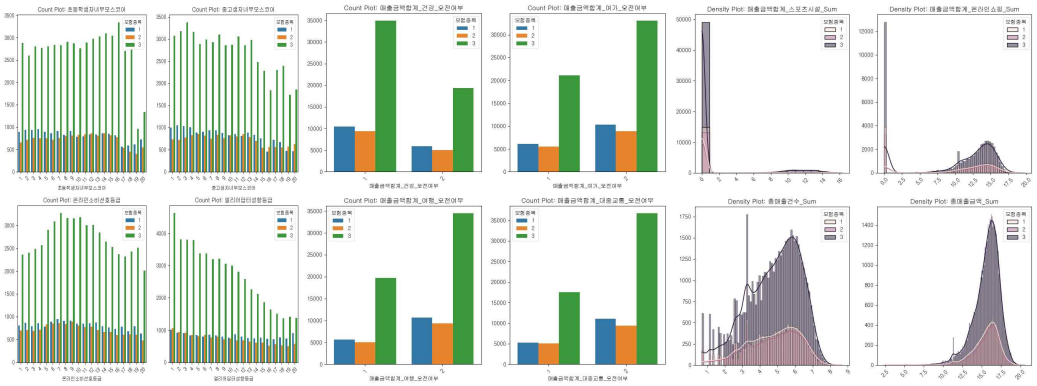


- 한편 전반적인 매출 관련 연속형 변수의 경우, MZ세대와 기성세대는 유사한 분포를 보였다.



(2) 보험종목에 따른 분석

- 각 소비자들이 가입한 보험의 종목에 따른 분석을 진행한 결과 전반적으로 유사한 소비 패턴을 보이는 것을 파악할 수 있었다. 이는 아래의 그래프를 통해 확인할 수 있다.



(3) 위의 과정을 통해, 기존의 연령 및 보험 종목에 따라서는 소비자의 일관된 지출 특성을 파악하는데 한계가 있음을 깨닫게 되었다. 이에 매출패턴을 보다 명확하게 파악할 필요성을 느끼게 되었으며, 매출 변수를 바탕으로 한 클러스터링을 통해 이를 해결하고자 했다.

3. 클러스터링

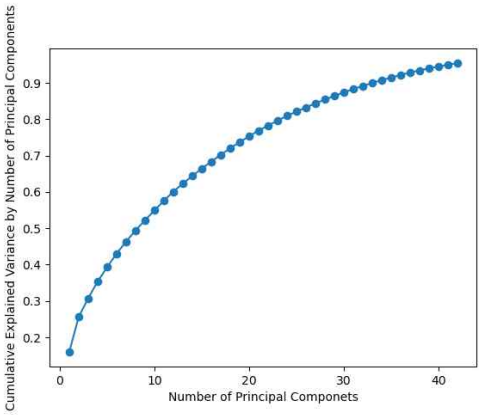
- 데이터 스케일링

· 스케일링을 진행한 이유: PCA 기법을 적용하기 전에 스케일링을 하지 않으면 변인이 가진 값의 크기에 따라서 설명 가능한 분산량이 왜곡될 수 있어 스케일링을 진행했다.

- PCA

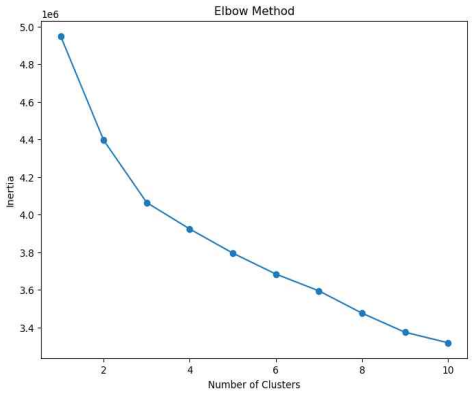
· 해당 기법을 사용한 이유: 사용 데이터는 카드 매출데이터로, 컬럼 수가 '61개'로 굉장히 많았다. 이 경우 컬럼이 너무 많아서 차원의 저주[1]가 발생해 클러스터링이 제대로 되지 않을 수 있어 먼저 PCA로 차원을 축소한 뒤에 클러스터링을 시도했다.

· PCA 주성분의 개수를 42개로 설정한 이유: 아래 그래프에서 주성분이 42개일 때 누적분산비율이 95% 이상이 되는 것을 확인하여, 데이터의 차원을 축소해 정보량의 손실이 적게 발생할 수 있도록 했다.



- Elbow Method

· 사용하려는 Kmeans 클러스터링 기법은 비지도 학습이지만 군집화하려는 클러스터의 개수를 지정해야 한다. 이때 Elbow method[2]을 활용해 적정 클러스터의 개수를 선정할 수 있다. 아래의 Elbow curve에서 기울기가 급격히 변하는 지점의 군집 개수가 3이었기에 군집을 3개로 지정해 클러스터링을 진행했다.



1) 차원의 저주: 차원이 증가하면서 학습데이터 수가 차원 수보다 적어져서 성능이 저하되는 현상
2) Elbow method: 군집수에 따라 군집내 총 제곱합 플롯팅하여 팔꿈치의 위치를 일반적으로 적절한 군집수로 선택하는 방법

- Kmeans 클러스터링

  · Kmeans 클러스터링 알고리즘 선택 이유: DBSCAN 방법과 비교했으나 해당 기법을 적용했을 때 군집화가 제대로 진행되지 않았다. 이는 클러스터링 대상 데이터가 고차원적인 데이터인데, DBSCAN은 고차원적인 데이터를 군집화하는 데에 강건하지 않은 기법이기 때문이라고 판단했다. Hierarchical 클러스터링은 연산시간이 크다는 특징이 있는데 컴퓨팅파워의 한계로 인해 적합하지 않다고 판단했다.

4. 클러스터링 결과 분석

- 클러스터 0에는 저연령대가 다수 포함되어 있고, 세 클러스터 중 MZ세대가 가장 많다. 추정 라이프스테이지를 보면 '초등자녀부모'가 압도적으로 많다. 클러스터 0과 1의 소비패턴을 비교해보면, 클러스터 1에 비해 '온라인쇼핑 매출금액 합계'의 평균이 높고, '요식 매출금액 합계'의 평균이 높다. 그리고 '쇼핑 매출금액 합계', '생활 매출금액 합계', '교육 매출금액 합계'의 평균 역시 높다.
- 클러스터 1에는 고연령대가 다수 포함되어 있고, 세 클러스터 중 MZ세대가 가장 적다. 추정 라이프스테이지를 보면 '성인 자녀 부모'가 압도적으로 많다. 클러스터 0과 1의 소비패턴을 비교해보면, 클러스터 0에 비해 '건강보조식품 매출금액 합계', '건강 매출금액 합계', '양약국 매출금액 합계'의 평균이 높다.
- 클러스터 2에는 연령대가 고르게 분포한다. 또한 소비수준이 높은 사람들이 많으며 고소득자의 비율이 높다. 추정 라이프스테이지는 고르게 분포하고 있다. 다른 두 클러스터에 비해 '반려동물보유추정등급'은 특별히 높은 수준으로, 반려동물을 보유하고 있지 않을 가능성이 높다고 볼 수 있다. '해외여행관심성향등급'은 클러스터 2가 특별히 낮은 수준으로, 해외여행에 대한 관심성향이 높다고 볼 수 있다.

5. 자동차 사고율 예측을 통한 결측치 처리

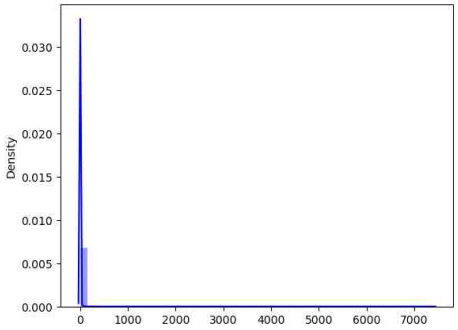
- 자동차보험의 보험료가 전부 결측치로, 사고율(손해액÷보험료)을 계산할 수 없어 사고율 자체를 결측치로 보아 대체했으며, 머신러닝 모델 LightGBM을 이용해 사고율을 예측했다. 자동차보험을 제외한 생명보험과 장기보험 사고율은 결측치가 없어 이를 이용해 모델을 학습시키고자 했다. 이때, 같은 손해보험으로 분류되는 자동차보험과 사고율 계산식이 동일한 장기보험 사고율로 학습을 진행했다. 또한 예측 성능을 최적화하고자 gridsearchCV를 활용하여 하이퍼파라미터 튜닝을 진행했다. 최적 하이퍼파라미터는 다음과 같이 도출되었다.

LGBMRegressor(max_depth=4, n_estimators=50)

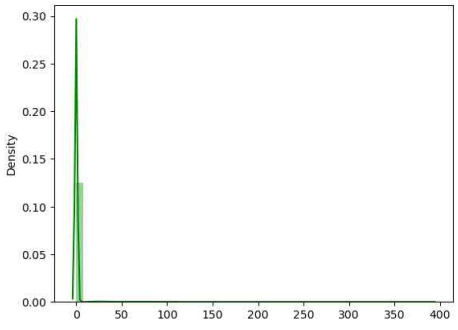
- 예측이 잘 되었는지 확인하는 성능평가지표로는 분산과 편향을 동시에 반영하는 RMSE를 선택했다. RMSE 값이 2.682로, GLM(General Linear Model)을 이용해서 예측한 사고율의 RMSE 8.445보다 훨씬 좋은 성능을 보였기에 최종 예측값으로 LightGBM을 이용한 결과를 사용하였다. 이렇게 산출된 사고율로 자동차보험의 결측치를 대체하였다. 대체가 잘 되었는지 확인하기 위해 학습데이터인 장기보험 사고율의 분포와 예측된 자동차보험 사고율의 분포를 시각화하여 비교했다. 두 분포 모두 사고율이 0 근방에서 밀도가 매우 높고, 0에서 먼 값이 매우 적은 비슷한 모양

을 보여 잘 대체되었다고 판단했다. 또한 두 분포 모두 감마분포의 확률밀도함수와 유사한 형태를 띠며 향후 분석 시 사고율의 분포로 감마분포를 가정할 수 있는 근거가 마련되었다.

[장기보험 사고율과 대체 후 자동차보험료 사고율 분포 첨부]



장기보험 사고율 분포



자동차보험 사고율 분포

6. 보험종목별 사고율 분석

- 머신러닝 모델이나 인공신경망 등 인공지능 기법을 활용한 사고율 예측이 최근 각광받고 있다. 그러나 주제의 특성상 사고율 예측보다는 분석을 통한 인사이트 도출과 현업에서의 활용이 더 중요시된다고 판단하였다. 따라서, 통계적 모델링을 활용하여 각 변수가 보험종목별 사고율에 어떻게 영향을 미치는지를 개별적으로 파악하는 분석을 진행하였다.

1) 생명보험

· 생명보험의 사고율 계산식은 '사고건수 ÷ 경과계약건수'로, 그래프를 통해 전반적으로 감마 분포와 유사한 형태를 보이는 것을 확인했다.

· 다만, 사고건수가 0인 데이터가 지나치게 많아[3] 사고율을 직접적으로 추정하는 것에는 무리가 있다고 판단해, 사고건수와 경과계약건수를 추정하는 모델을 각각 계산해, 이를 통해 최종적인 사고율을 추정했다.

1.1) 사고건수 계산 모델링:

· 사고건수는 '해당 연도에 발생한 보험사고 발생건수의 합계'이기에, 0 이상의 정수값을 가진다. 이에 0 이상의 가산변수를 정의역으로 갖는 포아송분포와 음이항분포가 '사고건수' 변수에 가장 적합하다고 판단해, 이를 바탕으로 한 GLM 모델링을 진행했다. 포아송 회귀와 음이항 회귀를 번갈아 사용하며 더 낮은 AIC를 갖는 회귀식을 최종 모델로 선정하고자 했으나, 종속변수를 음이항분포로 가정한 경우 모델의 계수가 계산되지 않아 최종적으로 포아송 회귀식을 통한 모델링을 진행했다.

· 이때, 연속형 설명변수의 경우 일부 변수의 계수가 발산해 모든 연속형 설명변수에 로그 변환을 진행했다. 이를 통한 최종적인 계산식은 아래와 같다.

3) 총 16377개의 데이터 중, 17건(약 0.1%)의 데이터만이 1건 이상의 사고건수를 기록하고 있었다.

사고건수에 대한 포아송 회귀모형

$$\log(\text{사고건수}) = \beta_0 + \beta_{\text{연속형}}\log(X_{\text{연속형}}) + \beta_{\text{범주형}}X_{\text{범주형}}$$

· 모델링은 전체 설명변수를 사용한 full 모델에서 p-value가 높은 순으로 유의하지 않은 변수를 제거해나가는 방식으로 변수 선택이 진행되었으며, 이를 통해 최종적으로 다음의 14개의 변수가 선택되었다.

· 다만, 클러스터 정보를 포함하는 'labels' 변수는 사고율에 고객 특성을 반영하고 있어 사고율과의 영향력을 직접적으로 추정하기 위해 변수에 포함하여 분석을 진행했다.

변수명	영향력	p-value
Intercept	-15.83070230	0.000279
매출건수합계_교육_Sum	-0.21086938	0.138605
매출건수합계_건강_Sum	-3.26670323	0.001550
매출건수합계_주차_Sum	-2.68380317	0.003899
매출건수합계_병원_Sum	2.04918221	0.011128
매출건수합계_건강보조_Sum	3.63591651	0.002504
매출건수합계_양약국_Sum	-2.03657900	0.055573
매출금액합계_자동차_Sum	-0.11914215	0.056797
매출금액합계_건강_Sum	1.22613683	0.000736
매출금액합계_주차_Sum	1.17865675	0.001986
매출금액합계_주유_Sum	0.11707220	0.062257
매출금액합계_병원_Sum	-0.74286664	0.011852
매출금액합계_건강보조_Sum	-1.28700274	0.007160
매출금액합계_양약국_Sum	0.76202141	0.092088
* labels1	0.06320788	0.902620
* labels2	-14.56935383	0.985047

* 범주형 변수

· 모델의 AIC는 287.69로 full 모델의 317.84에 비해 유의미하게 감소한 것을 확인할 수 있었다. 따라서 본 모델을 사고건수 계산에 대한 최적의 모델로 판단했다.

1.2) 경과계약건수 계산 모델링

· 경과계약건수는 '해당 자료년도에 보험계약의 효력이 유지된 기간을 고려한 계약 건수의 합계'이기에, 0 이상의 값을 가진다. 따라서 포아송 분포와 음이항분포가 해당 변수를 가장 잘 설명하는 분포라고 판단해, 이를 바탕으로 한 GLM 모델링을 진행했다. 사고건수와 마찬가지로, 모델링은 포아송 회귀와 음이항 회귀를 번갈아 사용하며 더 낮은 AIC를 갖는 회귀식을 최종 모델로 선정하고자 했으나, 종속변수를 포아송 분포로 가정한 경우 모델의 일부 계수가 발산해 최종적으로 음이항 회귀식을 통한 모델링을 진행했다. 사고건수에 대한 모델링과 마찬가지로 모든 연속형 설명변수에 로그 변환을 진행했다. '경과계약건수'에 대한 최종 모형은 아래와 같다.

경과계약건수에 대한 음이항 회귀모형

$$\log(\text{경과계약건수}) = \beta_0 + \beta_{\text{연속형}}\log(X_{\text{연속형}}) + \beta_{\text{범주형}}X_{\text{범주형}}$$

· 이때 음이항 회귀모형과 포아송 회귀모형 모두 로그함수를 연결함수(link function)로 가지는 GLM이기 때문에, 사고건수의 추정 회귀식과 동일한 형태의 계산식이 나타났다.

· 사고건수추정 모델과 동일하게 전체 설명변수를 사용한 full 모델에서 p-value가 높은 순으로 유의하지 않은 변수를 제거해나가는 방식으로 변수 선택이 진행되었으며, 클러스터 정보를 포함하는 'labels' 변수는 사고율에 고객 특성을 반영하고 있어 사고율과의 영향력을 직접적으로 추정하기 위해 변수에 포함하여 분석을 진행했다.

· 이를 통해 최종적으로 다음의 12개 변수가 선택되었다.

변수명	영향력	p-value
Intercept	-0.119819052	0.512514
* 성별2	0.052862245	0.000243
배달식품관심성향등급	0.059582542	2.57e-05
매출건수합계_요식_Sum	0.004629267	0.063586
매출건수합계_교육_Sum	0.004251200	0.051427
매출건수합계_스포츠시설_Sum	-0.046313136	0.096538.
매출건수합계_요양원_Sum	-0.017930694	0.230576
매출금액합계_자동차_Sum	-0.001283263	0.235696
매출금액합계_주차_Sum	0.002097015	0.077571
매출금액합계_스포츠시설_Sum	0.018645328	0.093538
매출금액합계_양약국_Sum	-0.001402588	0.187357
총매출금액_Sum	-0.006160305	0.305494
* labels1	-0.014890950	0.459941
* labels2	0.004515172	0.864716

* 범주형 변수

· 모델의 AIC는 39725로 full 모델의 Inf에 비해 상당 부분 감소한 것을 확인할 수 있었다. 이에, 본 모델이 최적의 모델임을 확인할 수 있었다.

1.3) 사고율 계산 모델링

· 앞서 추정한 사고건수와 경과계약건수 모델링을 통해, 사고율을 추정할 수 있었다. 앞의 과정과 마찬가지로 사고율 역시 로그 변환 후 모델링을 진행했다. 이에 대한 계산식은 다음과 같다.

최종 사고율 추정모형

$$\log(\text{사고율}) = \log(\text{사고건수}) - \log(\text{경과계약건수})$$

$$= \beta_{0_{\text{사고건수}}} - \beta_{0_{\text{경과계약건수}}} + \beta_{\text{연속형}_{\text{사고건수}}} \log(X_{\text{연속형}_{\text{사고건수}}}) - \beta_{\text{연속형}_{\text{경과계약건수}}} \log(X_{\text{연속형}_{\text{경}}}$$

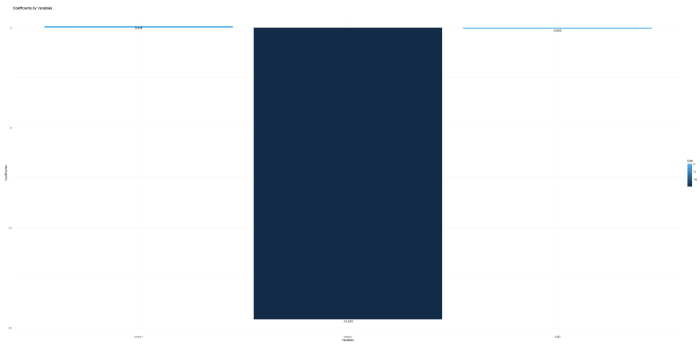
$$+ \beta_{\text{범주형}_{\text{사고건수}}} X_{\text{범주형}_{\text{사고건수}}} - \beta_{\text{범주형}_{\text{경과계약건수}}} X_{\text{범주형}_{\text{경과계약건수}}}$$

· 위의 식을 통해 최종적으로 다음 21개의 변수가 선택되었다.

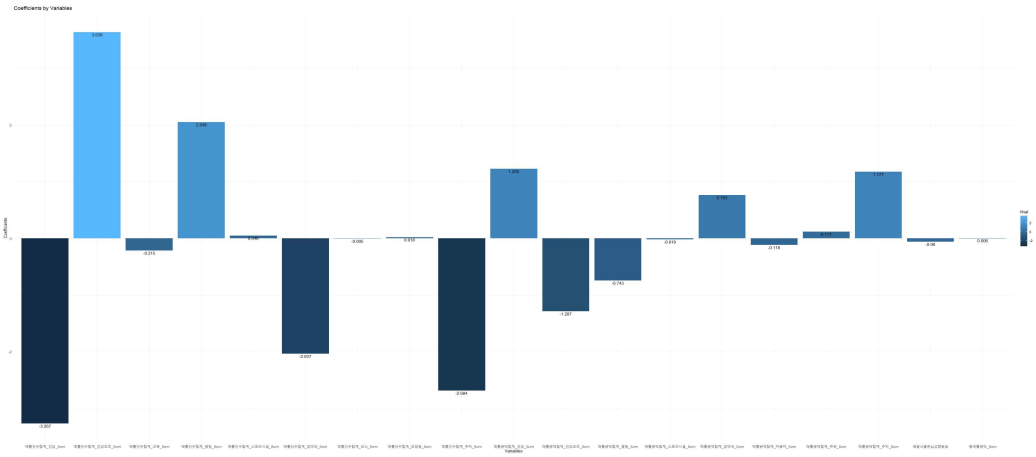
변수명	영향력
Intercept	-15.710883246
* labels1	0.078098826
* labels2	-14.573869005
매출건수합계_건강_Sum	-3.266703227
매출건수합계_건강보조_Sum	3.635916505
매출건수합계_교육_Sum	-0.215120580
매출건수합계_병원_Sum	2.049182208
매출건수합계_스포츠시설_Sum	0.046313136
매출건수합계_양약국_Sum	-2.036579002
매출건수합계_요식_Sum	-0.004629267
매출건수합계_요양원_Sum	0.017930694
매출건수합계_주차_Sum	-2.683803168
매출금액합계_건강_Sum	1.226136827
매출금액합계_건강보조_Sum	-1.287002741
매출금액합계_병원_Sum	-0.742866637
매출금액합계_스포츠시설_Sum	-0.018645328
매출금액합계_양약국_Sum	0.763424001
매출금액합계_자동차_Sum	-0.117858889
매출금액합계_주유_Sum	0.117072205
매출금액합계_주차_Sum	1.176559737
배달식품관심성향등급	-0.059582542
* 성별2	-0.052862245
총매출금액_Sum	0.006160305

* 범주형 변수

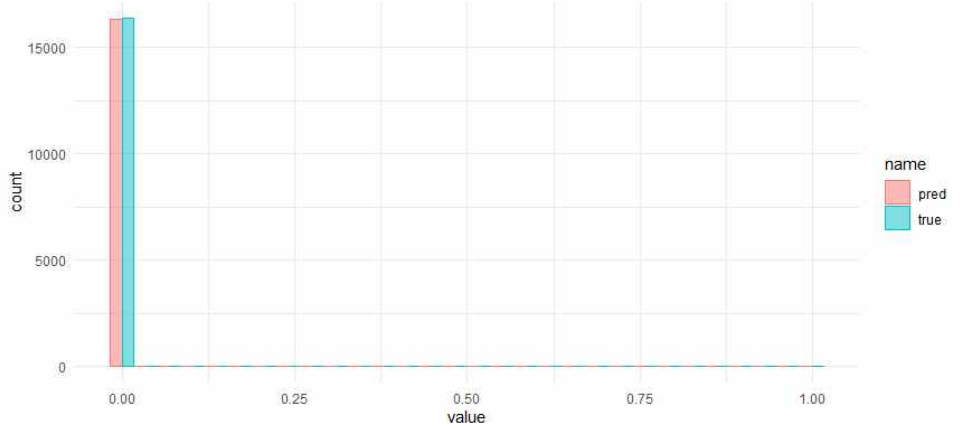
최종 사고율 추정 모델 변수 영향력 (범주형)



최종 사고율 추정 모델 변수 영향력 (연속형)



· 위의 과정을 통해 추론한 모델의 MSE(Mean Squared Error)는 약 0.0003으로 도출되었으며, 아래 그래프를 통해 추정 사고율과 실제 사고율의 분포가 매우 유사한 것을 확인했다.



· 이를 통해, 대다수의 소비 변수가 생명보험의 사고율 분석에 있어 유의미한 영향력을 행사했으며 이를 통해 추론한 모형의 정확도가 높은 것을 파악했다.

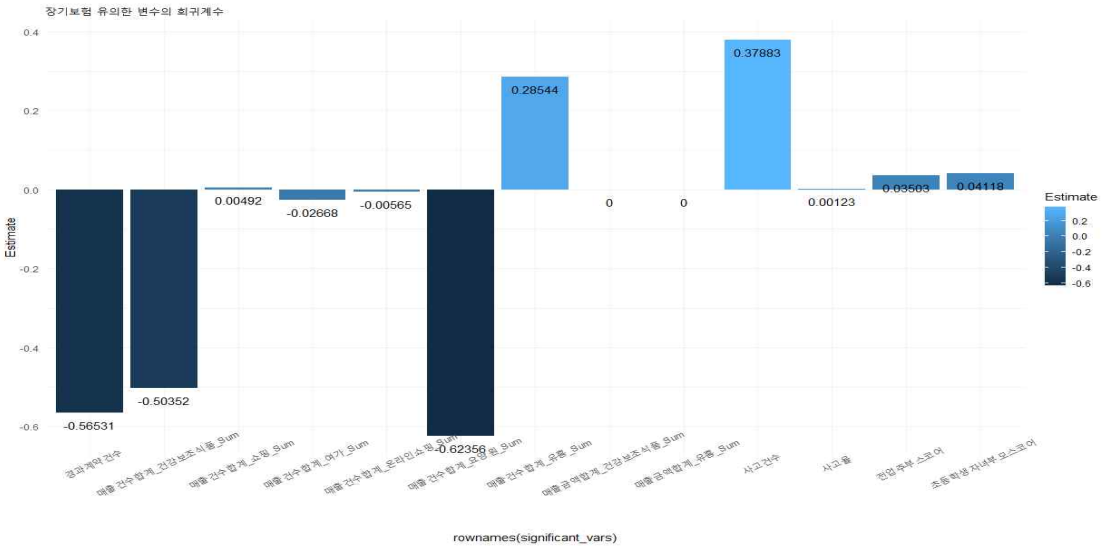
2) 장기보험

· 보험업종 사고율이 감마분포를 따른다는 선행연구를 참고하여 GLM 구축 시 사고율이 0인 관측치를 제거하여 감마분포와 구간을 일치시켰다. 또한 사고율의 단위가 0에 근접한 수부터 몇만까지 다양하게 이루어져 있기 때문에 로그를 취하여 scaling을 해주었다. 이렇게 생성한 로그감마모형에 VIF가 너무 높은 변수는 제거한 후 소비 관련 변수를 중심으로 모델 식을 세웠다.

장기보험 사고율에 대한 로그감마모형

$$\log(\text{사고율}) = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p$$

변수명	영향력	p-value
경과계약건수	-5.653e-01	0.00616
사고건수	3.788e-01	0.00233
추정_직업군1	1.526e+00	0.07573
추정_직업군2	1.574e+00	0.05588
추정_직업군3	1.835e+00	0.03650
추정_직업군4	1.565e+00	0.09456
추정_직업군5	1.388e+00	0.09628
초등학생자녀부모스코어	4.118e-02	0.01385
전업주부스코어	3.503e-02	0.09912
매출건수합계_여가_Sum	-2.668e-02	0.00386
매출건수합계_유흥_Sum	2.854e-01	0.04374
매출건수합계_쇼핑_Sum	4.922e-03	0.01096
매출건수합계_온라인쇼핑_Sum	-5.652e-03	0.04224
매출건수합계_건강보조식품_Sum	-5.035e-01	0.01040
매출건수합계_요양원_Sum	-6.236e-01	0.08323
매출금액합계_유흥_Sum	-2.594e-06	0.00631
매출금액합계_건강보조식품_Sum	2.022e-06	0.01390



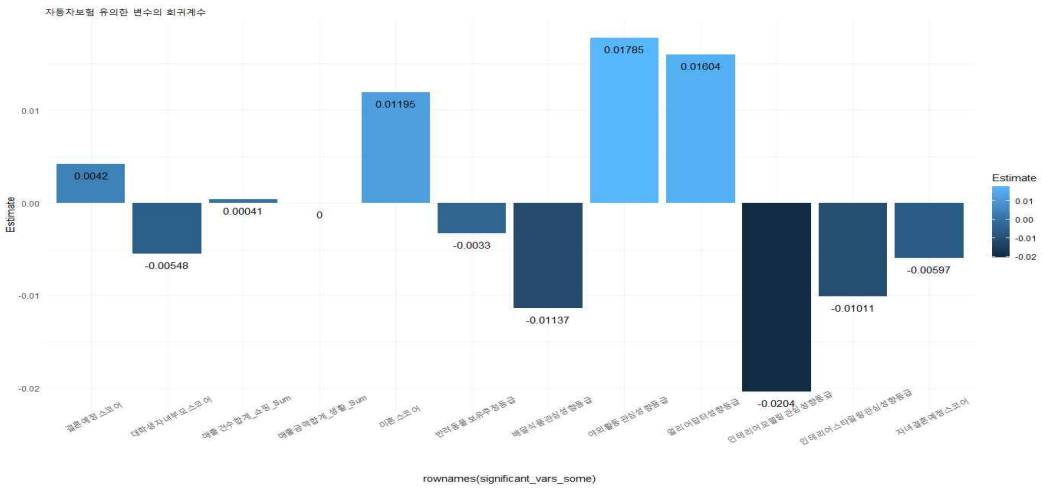
3) 자동차보험

- 자동차보험의 사고율은 장기보험 사고율을 기반으로 예측된 값이기 때문에 장기보험과 같은 이유로 로그감마모형을 사용했다. 장기보험과 마찬가지로 VIF가 너무 높은 변수를 제거했지만, 두개 뿐인 자동차관련 변수로 '자동차국외산구분코드'와 '자동차보험가입경력'를 포함하여 모델 식을 세웠다는 점에서 차이가 있다.

자동차보험 사고율에 대한 로그감마모형

$$\log(\text{사고율}) = \alpha + \beta_1 x_1 + \ldots + \beta_p x_p$$

변수명	영향력	p-value
사고건수	1.387e+01	< 2e-16
미혼스코어	1.222e-02	0.000169
결혼예정스코어	4.068e-03	0.086951
대학생자녀부모스코어	-5.617e-03	0.076196
자녀결혼예정스코어	-6.141e-03	0.019094
온라인소비선호등급	4.690e-03	0.083029
얼리어답터성향등급	1.624e-02	2.93e-10
프리미어성향등급	-3.981e-03	0.083659
배달식품관심성향등급	-1.154e-02	0.001604
반려동물보유추정등급	-3.331e-03	0.052384
야외활동관심성향등급	1.802e-02	2.28e-12
인테리어스타일링관심성향등급	-1.011e-02	0.000764
인테리어모델링관심성향등급	-2.040e-02	2.12e-14
매출금액합계_생활_Sum	3.860e-09	0.087727



7. 소비패턴에 따른 보험종목별 사고율 분석

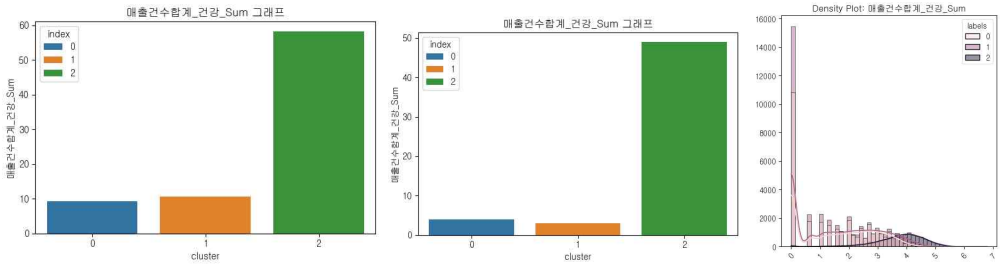
1) 생명보험

- 생명보험의 경우, 다양한 소비데이터와 사고율이 통계적으로 유의미한 관계성을 보이는 것으로 나타났다. 이때, 본 모형의 경우 연속형 변수는 설명변수와 반응변수가 모두 로그 변환이 적용된 로그-로그 모델의 형태를 지니고 있기에, 그 해석에 있어 직접적인 선형 관계성이 적용되지 않으므로 반응변수에 대한 변화율의 크기를 기준으로 해석이 진행되었음을 밝힌다.
- 우선, 남성이 여성에 비해 약 1.05배 사고율이 높은 경향을 보였다. 소비변수의 경우, 전반적으로 건강과 관련된 변수들이 생명보험의 사고율에 유의한 영향력을 미치는 것으로 드러났으며, 주로 매출건수와 관련된 변수가 매출금액과 관련된 변수에 비해 사고율에 더 큰 영향력을 미치는 것으로 드러났다. 아래는 사고율에 대한 변수의 영향력을 정리한 표이다.

사고율을 높이는 요인(+)	사고율을 낮추는 요인(-)
매출건수합계_건강보조_Sum(3.64%)	매출건수합계_건강_Sum(-3.27%)
매출건수합계_병원_Sum(2.05%)	매출건수합계_주차_Sum(-2.68%)
매출금액합계_건강_Sum(1.23%)	매출건수합계_양약국_Sum(-2.04%)
매출금액합계_주차_Sum(1.18%)	매출금액합계_건강보조_Sum(-1.29%)
매출금액합계_양약국_Sum(0.76%)	

- 군집별로 살펴보면, 클러스터 0의 경우 클러스터 1에 비해 사고율이 상대적으로 낮은 경향성(약 0.95배)을 보였다.
- 위에서 도출한 사고율 추정 모형은 이를 건강보조 항목의 매출건수로 인한 것으로 해석한다. 건강보조 매출건수 변수는 해당 변수가 1% 증가할 때 사고율이 약 3.67% 증가하는 방향으로 영향을 미치며, 해당 변수에 대해 매출건수가 더 많은 클러스터 1의 특성 상 클러스터 0에 비해 사고율이 더 높게 나타나는 경향성을 보였다. 다만, 두 집단 간 해당 변수의 크기 차이가 큰 편이 아니었기에, 두 집단의 사고율이 대체로 유사한 형태였다.
- 한편, 전반적으로 사고율에서 큰 차이를 보이지 않았던 클러스터 0 및 클러스터 1과 달리, 클러스터 2는 나머지 두 군집에 비해 사고율이 매우 낮게 나타나는 경향

을 보였다. 사고율 분석 모형은 이를 건강 항목에 대한 매출건수의 영향으로 해석하고 있다. 실제 클러스터 2는 타 군집에 비해 건강 항목에 대한 매출건수가 매우 높게 나타나는 것을 아래의 그래프들을 통해 확인할 수 있다.



- 매출건수_건강 항목은 변수가 1% 증가함에 따라 사고율이 약 3.27%씩 감소하는 것으로 드러났기 때문에, 클러스터 간 건강과 관련된 매출건수의 차이가 클수록 사고율은 기하급수적으로 감소하는 경향이 나타난다. 이와 같은 영향으로, 클러스터 2가 타 집단에 비해 매우 낮은 사고율을 보이는 것으로 이해할 수 있다.

2) 장기보험

- 장기보험 역시 사고율이 다양한 소비데이터와 통계적으로 유의미한 관계성을 보이는 것으로 나타났다. 이때, 반응변수에 로그 변환이 적용된 로그-감마모형이기 때문에, 그 해석에 있어 '독립변수가 한 단위 증가함에 따라 종속변수가 각자의 계수에 exponential을 취한 값만큼 증가한다/감소한다' 라고 선형적으로 해석했음을 밝힌다. 예를 들어, 매출건수합계_건강보조식품_Sum의 회귀계수는 -0.50352로 도출되었는데, 이는 소비자가 건강보조식품을 한 단위 더 구매하면 사고율이 exp(-0.50352)=0.6043배가 되어 감소하는 경향을 보인다고 해석된다. 아래는 사고율에 대한 변수의 영향력을 정리한 표이다.

사고율을 높이는 요인(+)	사고율을 낮추는 요인(-)
매출건수합계_쇼핑(1.004934)	매출건수합계_건강보조식품(0.6044)
매출건수합계_유흥(1.3302)	매출건수합계_여가(0.9736)
매출금액합계_건강보조식품(1.00002)	매출건수합계_온라인쇼핑(0.9943)
전업주부스코어(1.035651)	매출건수합계_요양원(0.5360)
초등학생자녀부모스코어(1.04204)	매출금액합계_유흥(0.9999)

- 구체적으로, 클러스터 0의 경우 클러스터 1에 비해 사고율이 평균적으로 낮은 경향을 보였다. 앞에서 도출한 사고율 추정 모형은 이를 건강보조식품의 매출건수로 인한 것으로 설명한다. 건강보조식품 매출건수가 더 많은 클러스터 1는 클러스터 0에 비해 사고율이 더 높게 나타나는 경향성을 보였다.
- 클러스터 0과 클러스터 2를 비교했다. 클러스터 0은 클러스터 2에 비해 건강보조식품 매출건수의 평균이 훨씬 낮다. 이는 모형에서 사고율을 높이는 방향으로 작용하기 때문에 클러스터 0의 사고율이 클러스터 2에 비해 높게 나타난다고 해석할 수 있다.
- 클러스터 1은 온라인쇼핑 매출건수가 다른 두 군집에 비해 낮게 나타났다. GLM 결과를 보면 온라인쇼핑 매출건수가 증가하면 사고율이 감소한다. 따라서 온라인쇼핑 매출건수가 낮다는 점이 클러스터 1에 속한 사람들의 사고율을 가장 크게 만드는 데에 기여한다고 해석할 수 있다.
- 또한 클러스터 1은 전업주부스코어가 군집들 중 가장 높다. GLM 결과 전업주부

스코어가 증가하면 사고율이 증가한다고 나타났다. 따라서 클러스터 1에 속한 사람들의 사고율이 가장 높다고 해석할 수 있다. 이는 전업주부일수록 사고율이 높은 경향성을 나타내므로, 사고율 예측에서 전업주부 여부를 활용할 근거가 될 수 있다.

- 한편, 클러스터 2는 다른 두 군집에 비해 사고율이 눈에 띄게 낮다. GLM결과 '매출건수합계_온라인쇼핑_Sum', '매출금액합계_유흥_Sum' 등 매출 관련 변수들이 높을수록 사고율이 낮다는 결과가 도출되었다. 따라서 클러스터 2의 사고율이 낮은 것은 '매출건수합계_온라인쇼핑_Sum', '매출금액합계_유흥_Sum' 등의 변수가 높게 나타나기 때문으로 해석할 수 있다.

3) 자동차보험

- 자동차보험의 경우, 사고율이 삼성카드 제공 등급 변수와 통계적으로 유의미한 관계성을 보이는 것으로 나타났다. 이때, 사용하는 GLM이 동일하기 때문에 장기보험 사고율 모형과 같은 방식으로 해석했다. 아래는 사고율에 대한 변수의 영향력을 정리한 표이다.

사고율을 높이는 요인(+)	사고율을 낮추는 요인(-)
배달식품관심성향등급(0.9886)	야외활동관심성향등급(1.01801)
인테리어모델링관심성향등급(0.9798)	얼리어답터성향등급(1.0161)
인테리어스타일링관심성향등급(0.9899)	온라인소비선호성향등급(1.004701)
프리미어성향등급(0.9996)	
반려동물보유성향등급(0.9967)	

* 좌측 : 등급(1~20)이 1에 가까울수록 사고율이 높아진다고 해석
* 우측 : 등급(1~20)이 1에 가까울수록 사고율이 낮아진다고 해석

- 우선, 야외활동관심성향과 얼리어답터성향의 경우 GLM에서 사고율을 낮추는 요인으로 판명되었다. 실제로 군집들의 성향 순위도 2 > 0 > 1 순으로, 사고율 순위가 1 > 0 > 2인 것과 음의 관계가 잘 보인다.

- 클러스터 0의 경우, '매출건수합계_쇼핑_Sum'가 크게 나타났다. GLM 결과에 의하면 '매출건수합계_쇼핑_Sum'가 증가하면 사고율이 증가하므로, 클러스터 0에 속한 사람들, 특히 많은 비율을 차지하는 젊은 층의 높은 쇼핑 매출건수가 사고율 증가에 기여했다고 해석했다. 또한 클러스터 0은 배달식품관심성향에서 클러스터 2보다 높았다. GLM 결과 배달식품관심성향이 높아지면 사고율이 증가한다. 이와 같은 영향으로 클러스터 0이 클러스터 2보다 사고율이 높게 나타난다고 이해할 수 있다.

- 클러스터 2의 경우, 세 군집 중 반려동물보유성향이 가장 낮다. GLM을 통해 반려동물 보유 성향이 높을수록 사고율이 증가한다는 것을 파악했고, 이에 따라 반려동물보유성향은 클러스터 2의 사고율이 낮아지는 데 기여한다고 볼 수 있다.

○ 최종결과

1. 결과 요약

1) 카드 소비 데이터 군집분석

- 카드 소비 데이터를 군집분석한 결과 3개의 클러스터가 도출되었다. 클러스터 0의 경우 MZ세대가 가장 많고 연령대가 낮은 사람들이 많다는 특

징이 있었다. 그리고 클러스터 1의 경우 연령대가 비교적 높은 사람들이 많았고 MZ세대가 가장 적었다. 클러스터 2의 경우 연령대가 고르게 분포하고 있고 소비 수준이 세 클러스터 중 가장 높다.

2) 소비패턴에 따른 보험종목별 사고율 분석

- 클러스터링을 통해 선정한 세 군집 각각의 소비패턴과 사고율의 연관성을 살펴보았다.

- 생명보험의 경우 카드매출 지표의 다수가 통계적으로 유의했고, 특히 매출건수가 사고율에 매출금액보다 큰 영향을 미치는 경우가 많았다. 군집별로 특성도 뚜렷하게 나타나 해석이 가장 용이했다.

- 장기보험 또한 카드매출 지표의 다수가 통계적으로 유의했으며, 매출건수가 사고율에 매출금액보다 큰 영향을 미치는 경우가 많았다.

- 자동차보험의 경우 삼성카드에서 제공한 각종 등급 변수가 유의했다. 그리고 연령대가 낮은 사람들에서 주로 보이는 특징이 사고율을 낮추는 요인으로 많이 도출되었다.

2. 핵심 아이디어

1) 삼성카드

① 군집별 맞춤형 카드 혜택 필터링 시스템

- 아이디어 제안 배경

• 카드사의 이익이 감소함에 따라, 여러 카드사들이 카드 혜택을 축소하고, 할인 혜택이 큰 일부 상품을 없애고 있다. 2023년 7월 3일 금융업계에 따르면 8개의 카드사는 올해 상반기 159개 카드의 신규 가입을 중단했는데, 이는 지난해 연간 단종 카드 개수(116)를 훌쩍 뛰어넘는 수치이다. 신한카드에서는 '더모아카드'의 분할 결제 혜택을 축소 통보해 많은 소비자의 불만을 사기도 했다. 이러한 배경에는 카드사 수익성 악화가 존재한다. 삼성카드는 지난해 1분기보다 영업 이익이 약 11.4% 줄어들었고, 하나카드는 약 66.2% 감소했다. 혜택 축소는 이윤 창출이 목적인 기업 입장에서는 고려 대상이다. 하지만 카드 혜택을 재편할 때, 고객의 관심이 가장 적은 혜택을 취사 선택해 폐지하는 것이 고객 이탈을 막는 데에 보다 효과적일 것이다.

• 카드 상품 개발에 있어서도, 혜택을 무조건 많이 넣는 것보다 타겟 고객의 특성을 고려해 혜택을 추가하는 것이 비용적인 측면에서 합리적일 것이다. 이러한 측면에서 카드사 고객을 군집화한 뒤, 군집별 특성을 도출하고 해당 특성을 고려한 맞춤 혜택을 선택적으로 넣는 것이 적절할 것이다.

- 아이디어 세부 내용

• 삼성카드 소비데이터를 군집화한 결과를 활용해 '삼성 id vita 카드' 할인 혜택을 재구성해보았다. 클러스터 0의 경우는 '생활', '쇼핑' 등의 항목에서 많은 소비를 했다. 클러스터 1의 경우는 '의료', '건강', '양약국', ' 보험' 등의 항목에서 많은 소비를 했다. 클러스터 2의 경우는 모든 항목에서 세 클러스터 중 가장 높은 소비금액을 기록했다. 삼성 id vita 카드가 병원비, 보험료 등 의료비용에 지출금액이 많은 고객을 타겟으로 하기에, 클러스터 1이 해당 카드의 타겟 고객군으로 가장 적절할 것이

다. 분석 결과에 따르면 해당 고객군에 삼성 id vita 카드의 '의료/보험 영역' 할인 혜택은 매우 적합하다. 건강보조식품에 대한 소비도 많기에 '헬스/뷰티 영역'의 혜택도 대체적으로 적절하다. 하지만, 해당 클러스터의 경우 온라인 소비 성향이 적기 때문에 헬스/뷰티 항목 세부에서 '아모레몰' 할인과 같은 온라인몰 혜택은 적절치 않을 수 있다. 전반적인 생활 영역 소비 수준은 클러스터 1이 가장 적기에 '이동통신/렌탈/멤버십 영역'의 할인 혜택도 축소하는 방향이 적절할 것이라고 분석했다.

- 아이디어 기대효과

· 고객의 특성과 관련 있는 혜택들만을 담아 만든 카드를 통해 소비자의 마음을 사로잡을 수 있다.

· 카드사 입장에서는 할인 혜택을 필터링하는 것을 통해 각종 비용(혜택 문의 비용, 할인 혜택 제공을 위한 제휴 비용)을 절감할 수 있다.

2) 보험개발원

① 카드 소비 변수 활용 보험 손해율 예측 모형 고도화 방안

- 아이디어 제안 배경

· 최근 대형 손해보험사들의 손해율이 증가하면서 재무상태에 위험이 발생하고 있다. DB손보의 경우 자동차보험 손해율이 77.0%, 현대해상의 경우 77.5%를 기록하며 주요 보험사 모두 전년 동기에 비해 손해율이 상승하는 추세였다.

회사명	2022년 6월 자동차보험 손해율	2023년 6월 자동차보험 손해율
현대해상	75.7	77.5 (▲)
KB손해보험	75.0	77.7 (▲)
DB손해보험	71.9	77.0 (▲)
AXA손해보험	85.4	89.1 (▲)
하나손해보험	85.6	95.5 (▲)

· 자동차보험 손익분기점이 손해율 78~80%인 점을 고려하면 매우 높은 수준이다. 올 상반기 12개 손해보험사의 자동차보험 손해율은 평균 89.2%로 집계되었다. 소형 보험사들은 손해율이 90%에 육박할 정도로 어려움을 겪고 있다. 손해율 상승이 지속되면 초반에는 보험료 인하의 가능성도 있지만, 장기적으로는 소비자에게 비용이 전가되는 결과를 낳을 수 있어 고객 이탈률이 상승하는 악순환이 발생할 수 있다. 손해율을 사전에 더 정교히 예측할 수 있다면 보험계약자에 대해 보다 엄밀한 판단이 가능해져 손해율을 감소시킬 수 있기에, 손해율 예측은 보험사와 고객 모두에게 중요한 부분이다. 따라서 본 팀은 분석에서 유의미했던 카드 소비 관련 변수를 활용해 손해율 예측 모형 고도화 방안을 제시하고자 한다.

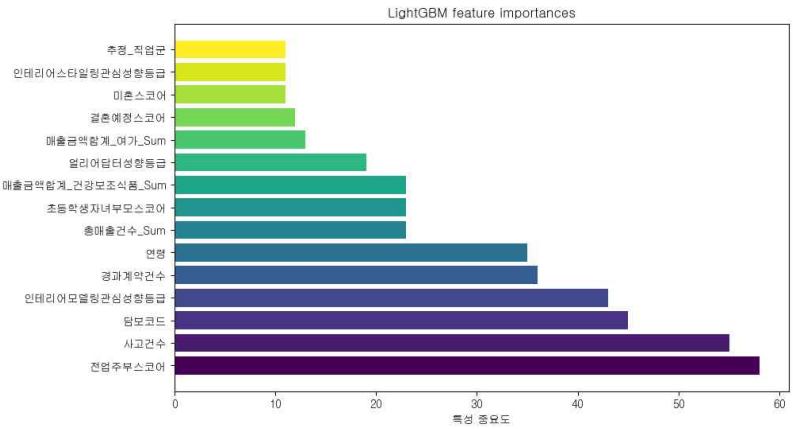
- 아이디어 세부 내용

· 아래 표는 GLM 결과를 분석했을 때 도출된 '사고율에 유의한 영향을 미치는 카드 소비 관련 변수'이다. 아래와 같은 변수를 보험 종목별 사고율 예측에 활용할 것을 제시한다. 이와 같은 카드 소비 데이터를 활용한

사고율 예측 결과를 언더라이팅 시 보조적 지표로 활용할 수 있을 것이다.

보험	방향성	변수명
생명보험	양의 방향 (+)	건강보조, 병원 등의 매출 건수 합계
	음의 방향 (-)	건강, 주차 등의 매출 건수 합계
장기보험	양의 방향 (+)	쇼핑, 유흥 등의 매출 건수 합계
	음의 방향 (-)	건강보조식품, 여가 등의 매출건수합계
자동차보험	양의 방향 (+)	배달식품관심성향, 프리미어성향
	음의 방향 (-)	야외활동관심성향, 얼리어답터성향

· 또한 본 대회에서 사용한 통계 모형인 GLM을 통해서도 사고율을 예측할 수 있지만, 실제 더 많은 고객 데이터로 사고율 예측을 진행한다면 LightGBM과 같은 머신러닝 기반 모형을 사용할 것을 제안한다. 머신러닝 기반 모형을 사용할 시에, Feature Importance와 같은 XAI(설명가능한 인공지능) 기법을 이용할 것을 제안한다. 아래 그래프는 본 대회에서 제공한 삼성카드 매출 데이터와 보험개발원 데이터를 활용해 장기보험 손해율에 대한 LightGBM 활용 예측을 진행한 후, Feature Importance를 도출한 것이다. 이와 같이 머신러닝 모델링 결과에 대한 근거를 제시하는 것을 통해 머신러닝 모델의 신뢰도를 제고할 수 있다.



- 아이디어 기대효과

· 언더라이팅 심사 지표를 카드 소비 데이터로 확대함으로써 엄밀한 심사를 할 수 있어 보험사 입장에서는 손해율 완화에 도움이 될 수 있다. 뿐만 아니라 보험 계약을 승인 받는 소비자 역시 손해율에 대해 음의 영향을 미치는 카드 소비 변수를 통해 보험료 할인을 적용받게 된다면 만족감이 높아질 것이다.

· 통계적 모델과 머신러닝을 활용해 손해율 예측 모형을 고도화하게 된다면 방대한 양의 고객 데이터를 대상으로 한 계약 심사의 정확성과 속도를 모두 제고할 수 있을 것이다.

② 클러스터 특성을 고려한 맞춤 보험 상품 추천

- 아이디어 제안 배경

• 보험개발원은 최근 보유 보험 정보와 타 기관의 금융/비금융 데이터를 결합해 인사이트를 도출하고자 한다. 따라서 본 팀은 분석 결과를 보험개발원 측에서 활용 가능한 형태로 제시하고자 삼성카드의 카드소비 데이터 분석을 통해 얻은 클러스터별 특성을 보험상품과 연결지어 클러스터별 추천 보험상품을 도출했다.

- 아이디어 세부 내용

• 클러스터 0의 경우는, MZ세대의 비율이 높다. 따라서 해당 클러스터에 속한 고객에게는 가성비와 보험 가입의 편의성을 중시하는 MZ세대의 특성을 고려해 보험기간이 짧고 1만원 이하의 소액상품인 미니보험상품을 추천할 수 있다. 그리고 해당 클러스터가 여행 관련 소비가 많다는 점을 고려해 '온라인 해외여행자보험'을 추천할 수 있다.

• 클러스터 1의 경우는, 고령자 비율이 다른 클러스터에 비해 높다. 그리고 건강, 양약국 등의 의료·건강 분야 관련 지출이 많기에, 유병자보험상품을 추천해줄 수 있다. 그리고 건강에 대한 소비가 많을 뿐 아니라, 건강관리 관심 성향이 높기에 건강을 잘 관리할수록 보험료 할인 혜택이 있는 상품을 추천할 수 있다.

• 클러스터 2의 경우는, 고소비 집단이고 프리미어 소비 성향이 높다. 그리고 전문직의 비율이 다른 클러스터에 비해 높은 편이다. 해당 집단은 소득이 많을 뿐 아니라 지출 역시 많기에 자산 관리를 체계적으로 할 수 있도록 도와주는 것이 필요하다고 본다. 따라서 '연금저축보험' 등의 자산관리상품을 추천할 수 있다.

- 아이디어 기대효과

• 보험가입에 대한 관심이 적은 MZ세대에게 적절한 보험상품을 추천해줌으로써 MZ세대의 보험 가입률 개선 효과를 불러올 수 있다.

• 추천 서비스를 통해 보험상품을 탐색하는 수고를 덜어줄 수 있어 고객 만족도 제고를 기대할 수 있다.