카드 데이터를 활용한
사고율 예측 개선 검증과 요인 분석 및
군집화를 통한 고객 특성 분류

Ⅰ. 개요

1. 요약

보험사는 ‘수지상등의 원칙’에 기반하여 보험료를 책정한다. 해당 원칙을 유지하기 위해서는 고객의 특성을 고려한 적합한 사고율의 파악이 필수적이다. 하지만 기존에는 보험료를 산출하느 과정에 있어서 사고율 산출 시, 보험에 가입할 때 고객이 기입하는 개인의 인적정보만을 이용해 판단할 수 밖에 없는 한계가 존재했다. 하지만 마이데이터 산업의 성장으로 사고율 산출과정에서 개인의 생활 전반에 대한 데이터를 확보할 수 있다면, 개인의 특성을 잘 반영할 수 있는 합리적, 객관적인 보험료를 산출할 수 있다. 본 분석에서는 마이데이터, 즉 대회에서 주어진 카드데이터를 기존 보험개발원 데이터에 추가하여 사고율 예측 모형을 구축할 때 성능의 변화를 파악할 것이다. 이를 통해 마이데이터를 사고율 예측에 함께 고려할 수 있다면 사고율 예측을 더 효과적으로 할 수 있는지 검증할 것이다. 또한, 앞선 검증으로 마이데이터를 활용한 모형이 더 우수하다고 판단되면, 보험료 초기 산정시 사용한 사고율(=보험개발원 데이터만을 이용해 산출한 사고율)과 카드데이터를 함께 사용한 사고율을 이용해 기존 고객을 4가지 분류로 군집화할 것이다. 이를 통해 보험사의 수익을 높이고, 손실을 방지하기 위해 각 고객의 유형별 전략을 생각해 볼 것이다. 또한, XAI(설명가능한 인공지능) 기법을 사용하여 각 모델에 대해서 사고율에 영향을 끼치는 요인을 분석하며, 합리적인 보험료 산출을 위해 보험사가 집중적으로 고려해야 할 요인을 파악하고자 한다.

2. 분석배경

보험사가 고객에 대한 보험료를 산출하는 과정에서 필히 지켜져야 하는 원칙이 있다. 이는 ‘수지상등 원칙(=등가원칙)으로 각 위험집단으로부터 납입되는 보험료의 총액이 위험집단에게 지급되는 보험금의 총액과 같아야 한다는 것이다. 따라서 위험집단, 즉 고객의 보험료를 산출하는 과정에서 고려해야하는 ‘사고율’ 예측은 매우 중요하다. 예측의 결과에 따라 심각한 적자가 생길 수도, 회사의 전체적인 수익을 극대화시킬 수 있기 때문에 고객의 특성에 따라 사고율을 정확하게 예측해야 하는 것이다.

사고율 예측은 단순히 손실을 방지하기 측면만 아니라 고객 관리의 영역에서도 필수적이라 할 수 있다. 복잡한 공식에 의해 산출된 보험료가 사고율을 바탕으로 어떤 근거로 책정되었는지 효과적으로 설명할 수 있다면, 고객은 보험상품과 보험료에 대해 더욱 신뢰하게 될 것이다. 이는 보험사의 입장에서 신규 고객의 유입을 유도할 수 있다. 또한, 기존고객을 대상으로 가입된 이후에 축적된 데이터를 기반해 사고율에 영향을 끼치는 요인을 분석할 수 있다면 보험료 인상/인하, 프로모션, 집중적 관리 등 고객의 이탈을 방지하고, 만족도를 높일 수 있는 다양한 전략을 세울 수 있다는 장점이 있다.

그러나 여러 보험사의 현존 사고율 예측 방식에는 한계가 분명히 존재한다. 초기 보험료 산정 시, 고객이 기입한 제한적인 신상정보에 의존하여 사고율을 산정하는 경우가 많다. 또한, 그 이후의 사고율 산정 공식 역시, “사고건수/계약건수”, “보험금/보험료” 등 다소 정적인 모델링 방식으로 시간에 따른 변화나 외부 요인의 영향을 반영하기 어렵다. 이는 사

고율에 영향을 줄 수 있는 고객의 새로운 패턴이나 동향을 제대로 파악하지 못한다. 일례로, 개인의 사고율 예측을 위해서는 생활, 소비 패턴 등 수 많은 요인이 존재하는데, 기존 방식은 나이, 성별, 보험가입경력 등 제한적인 신상 데이터만을 이용하여 판단해 그 정확도가 낮다.

또한, 사고율을 예측하는 모델의 성능이 뛰어나다고 하더라도, 몇몇 예측 모델은 그 구조가 매우 복잡하여 결과를 해석하는데 많은 어려움을 겪고 있어 보험사의 의사결정과정을 이해하기 원하는 고객의 수요와 상충되고 있다. 즉, 특정 고객의 사고율이 높게 측정되었다면 어떤 요인들로 인해 해당결과가 산출되있는지 설명할 수 있어야 좋은 사고율 예측모델이라고 할 수 있다.

이러한 점을 근거로, 보험사의 수익 극대화, 고객 유치 및 신뢰도 상승을 위해 목적에 맞게 세분화된 새로운 사고율 예측 기법의 도입이 필수적이다. 본 연구에서는 사고율을 정확하게 예측하는 모델뿐만 아니라, 고객에게 사고율에 대한 근거를 명확하게 하고 쉽게 설명할 수 있는 모델을 구축할 것이다. 그 과정에 있어서 AI와 같은 모델링 기법 역시 필요하지만, 대회에서 제공해준 카드 데이터, 즉 일명 '마이데이터'를 활용해 혁신적인 분석을 할 예정이다. 기존 보험개발원 데이터뿐만 아니라, 카드사 데이터를 함께 활용한다면 개별 고객의 생활 전반의 정보를 폭넓게 이해할 수 있게 되어 '개인화된' 예측을 할 수 있다. 시간에 따라 변화하는 고객의 행동을 반영할 수 있는 동적인 모델을 만들 수 있는 것이다. 해당 내용을 근거로, 분석 목적의 명확성을 위해 두 가지 가설을 설정하였다. 그 내용은 하단과 같다. 해당 가설을 검증하는 방식으로 심도 있는 분서글 진행하려고 한다.

[가설1]

사고율 예측과정에서 보험개발원 데이터에 카드데이터(마이데이터)가 추가된다면, 더 정확한 사고율 예측을 할 수 있을 것이다.

[가설2]

정확히 예측된 사고율 모형에 중요한 영향을 끼치는 요인(변수)과 그 영향력을 파악할 수 있다면 더 정확한 보험료를 산출할 수 있을 것이다.

3. 아이디어 제안

두 가지 가설에 대한 검증 전략은 다음과 같다.

첫 번째 가설의 경우, 보험명세서 데이터만을 이용한 사고율 모형(이하 "A" 모델)의 성능이 보험명세서 데이터와 카드데이터를 함께 이용한 사고율 모형(이하 "B" 모델)의 성능보다 낮음을 확인하며 검증할 수 있다. A모델의 성능이 B모델보다 좋지 않다면 보험명세서 데이터만을 통해 산출한 사고율이 부정확하다는 가능성을 내포하는 것이고, 이는 보험사의 손실을 유발할 수 있다고 유추할 수 있다.

다만, 해당 보험명세서 데이터의 경우 보험의 종류가 생명보험, 장기손해보험, 자동차보험으로 나누어 제시되어 있다. 이를 통합적으로 모델링 할 경우, 해당 보험의 분류 하위에 보장내용 역시 포함되어 있기에 모형의 해석에 있어 굉장히 어려울 수 있다. 따라서 해석의 용이함을 위해 3가지 보험 종류에 따라 각기 다른 사고율 예측 모형을 만들 것이다. 또한,

각 보험별 보장내용을 기반으로 총 15개의 모델(생명보험 3개, 장기손해보험 6개, 자동차보험 6개)을 생성하는 경우 역시 고려해보았지만, 데이터의 정량적인 한계로 인해 모델의 훈련이 Underfitting 될 가능성이 농후하기에 보험의 3가지 종류로 모형을 만들고 볼머별 각 보장내용을 따로 해석하려고 한다.

또한, 목차의 결과보고서 항목에 자세히 서술하겠지만, 해당 3가지 보험별로 제시된 사고율 공식에 따라 단순히 사고율을 종속변수로 설정하고 예측을 수행하면 데이터의 한계(NULL, 결측치, 0값)로 인해 모델의 학습이 제대로 이루어지지 않아 본래 의도한 검증을 잘 수행할 수 없다. 따라서 보험별 모델링의 과정에서 서로 다른 전략을 구축했다. 이는 다음과 같다.

[생명보험]

주어진 사고율 공식을 그대로 적용한 결과 총 16377개의 데이터 중 0이 아닌 값은 7개뿐이었다. A모델과 B모델을 비교하기에는 현실적으로 성능이 높은 모형을 만들 수 없다는 한계가 존재한다. 뒤에 서술될 장기손해보험이나 자동차보험처럼 데이터의 한계로 인해 새로운 사고율을 정의하는 방식 역시 존재하지만, 대회에서 주어진 사고율 공식을 따랐을 때, 대부분이 0 값일뿐 결측치는 단 10개였기에, 자의적으로 사고율을 정의하고 해석한다는 것은 근본적인 가정의 오류를 버할 수 있다고 판단했다. 따라서 'ACCD_CUNT' 변수를 활용하여 A,B 모델을 비교하는 것이 아닌, 사고가 일어나는데 영향을 끼치는 변수를 '해석'하는 측면에 더욱 집중하였다.

[장기손해보험 및 자동차보험]

대회에서 주어진 사고율 공식을 바탕으로 사고율을 구했을 때, 장기손해보험의 경우 0이 아닌 값은 14414개 중 365개, 자동차보험의 경우 54210개 중 모든 값이 결측치였다. 두 보험의 사고율이 동일하다는 것을 고려하여, 장기손해보험에 일부 0이 아닌 값의 데이터가 존재하더라도 기존 공식을 그대로 사용하여 두 모델에 적용하는 것은 결과적인 오류가 생길 수 있다. 따라서 새로운 사고율을 정의하는 방식이 필요하다고 판단했다. 'ACCD_CUNT'를 종속변수로 설정하여 Classification 모델을 구축한 후, 일반적인 Binary 형태의 예측값인 0,1을 사용하는 대신 해당 범주를 결정하는 예측확률값(사고가 일어날 확률)을 이용하여 해당값을 새로운 사고율로 대체하였다. 일반적인 분류 모델은 Cut-off 값을 0.5로 설정하고 예측확률값이 0.5 이상이면 1, 0.5 미만이면 0으로 분류하는데, 해당 원리에서 착안하였다. 자세한 설명은 목차의 '결과보고서' 부분에 서술할 것이다.

따라서 Classification을 통해 A,B 모델의 성능을 비교하여 어느 모델이 더 우수한 모델인지 파악할 것이다. 이를 통해 만약 마이데이터를 함께 활용한 B모델의 성능이 더 좋다면, 보험개발원 데이터만을 사용하여 모델링한 A모델의 사고율과 카드 데이터를 함께 사용하여 모델링한 B모델의 사고율을 바탕으로 군집화하여, 고객의 유형을 분류할 것이다. 이를 통해 보험에 가입하고 있는 기존 고객을 특성에 맞게 분류하여 보험사의 손해를 줄이기 위한 전략을 고민할 것이다. 이 과정의 경우 A모델을 사용한 사고율은 보험데이터만을 사용했기 때문에 보험사가 예상한 **예상사고율**, B모델을 사용한 사고율은 카드데이터를 함께 사용했기 때문에 **실제 사고율**이라는 가정을 내포한다.

두 번째 가설의 경우, 보험종류별 3가지 모델링을 한 후, 종속변수인 사고율에 영향을 끼치는 변수를 파악하기 위한 과정이다. 이는 신규 고객의 보험료 초기 산정시, 사고율에 중요한 영향을 끼치는 요인을 사전에 파악하여 올바른 보험료의 설정으로 보험사의 손실을 최대한 줄일 수 있다. 또한, 중요한 변수를 알 수 있다면 후에 현업에서 실제로 사고율을 모델링하는 과정에서 수많은 마이데이터 변수 중 중요한 변수를 집중적으로 선택하여 예측모형의 무게 역시 간소화할 수 있다는 장점이 있다. 또한, 초기 보험료 산정 시, 고객의 신상 조사에 있어서도 집중적으로 분석해야 할 요인 역시 확인할 수 있다.

[생명보험]

우선적으로 사고유무를 바탕으로 사고가 일어난 데이터(사고유무=1)가 총 17개가 존재한다는 점을 고려하여, 적은 Label(종속변수)을 가진 데이터의 개수에 맞춰 총 데이터를 감소시키는 Undersampling 기법을 활용하여 모델링을 진행할 것이다. 하지만 34(=17*2)개의 데이터를 통해 학습시킨 모델을 비교할 수 없기에, A와 B모델의 성능에 대한 분석을 제외하고 비교적 설명력과 모델의 높은 성능을 동시에 만족하는 모델을 사용하여 사고율에 영향을끼치는 요인을 해석할 것이다.

[장기손해보험 및 자동차보험]

앞서 설명한 것처럼 새롭게 사고율을 정의하는 Classifcation 방식을 통해 모델링을 완료한 후, A/B 모델에 대해서 변수의 영향력을 파악할 것이다. 이때, 설명가능한 인공지능(모형의 예측결과를 사람이 이해할 수 있고, 신뢰할 수 있게 하기 위한 기법의 총칭)의 기법 중 ICE(Individual Condition Expectation) 방법론을 사용하여 사고율에 영향을 끼치는 각 변수의 영향력을 분석할 것이다. 이때, 중요한 변수를 대상으로 그 영향력을 확인할 것인데, 중요도의 경우 앞선 Classification 모델링 과정에서 모델의 설명력이 높은 Tree 기반 모델을 사용하여 Feature Importance를 Scree Plot 그래프로 시각화한 후, 중요도가 갑자기 감소하는 부분을 기준으로 중요변수를 추출하려고 한다.

Ⅱ. 주요기술

본 프로젝트에서는 다양한 분석 기술의 접목을 시도하였다. 분석 기술들의 핵심 목적은 마이데이터의 카드 거래 내역을 기반으로 기존보다 사고율 예측이 뛰어난 모델을 만드는데 있었다. 또한 전반적인 예측 성능의 향상뿐만 아니라 모델의 “설명가능성”에도 집중하였다. 모델을 통해 비즈니스 관점에서 보험 분야의 맥락에 맞는 의사결정이 내려지고 고객들이 결과가 내려지게 된 과정을 논리적으로 이해할 수 있도록 해석력을 키우기 위해 다양한 전처리 및 머신러닝/데이터마이닝 모델링 기법을 시도하였다. 해당 섹션에서는 사용한 분석 알고리즘의 명칭 및 간단한 설명을 첨언하였고, 알고리즘이 구체적으로 쓰인 용도나 맥락 추후 보고서의 “데이터 분석 및 모델링” 부분에서 더 자세히 기술하였다.

(1) 예측 모델링 과정에서 활용한 분석 기술

- 로지스틱 회귀

이진 분류 문제에 사용되는 통계 기반의 머신러닝 알고리즘이다. 생명보험 데이터셋의 사고율을 예측하기 위해 활용되었다. 사고율을 예측하기 위한 변수로 accd_count를 정의하였다. 단 한번이라도 사고가 발생한 고객일 경우 1로, 그렇지 않은 무사고 고객일 경우 0으로 정의가 된다. 따라서 모델의 종속변수는 1과 0 중 하나를 예측하게 되는 이진 분류(Binary Classification)의 문제가 된다. 로지스틱 회귀의 특징은 선형 결정 경계를 찾아 데이터를 분류하고 이를 통해 어떤 클래스에 속하는지 확률적으로 예측한다는 점이다. 또한 시그모이드 함수를 사용하여 입력 데이터와 가중치의 선형 결합을 변환하는 것을 통해 확률값을 생성한다. 특정 threshold와 비교하여 0~1 사이의 확률값을 기반으로 클래스를 할당한다.

- Gradient Boosting

트리 기반 모델 중 하나로 앙상블 기법을 사용하여 높은 성능을 발휘한다. 기존의 약한 예측 모델들을 결합하여 더 강력한 예측 모델을 형성하는 방식으로 작동한다. 의사결정 트리를 기반으로 연속 학습을 통해 오차를 보정해나가는 방식으로 모델이 구축된다. 뿐만 아니라 고차원의 복잡한 데이터에도 잘 적응하며, feature scaling 등의 전처리 과정을 상대적으로 덜 필요로 한다. 그리고 결정적으로 변수의 중요도를 평가할 수 있는 기능을 제공하여 모델의 해석 가능성을 높여준다.

[하이퍼파라미터]

- n_estimators: 학습 과정의 반복 횟수. 클수록 모델의 복잡도가 증가하나 과적합 가능성도

커진다.
- learning_rate: 각 weak learner의 학습 기여도를 조절하는 파라미터다. 작을수록 학습 과정이 조심스럽게 이뤄지며 더 많은 반복을 필요로 한다
- max_depth: 각 의사결정 트리의 최대 길이를 나타낸다. 트리 길이가 길수록 더 복잡한 패턴을 학습할 수 있으나 마찬가지로 과적합 위험이 커진다.

- Extra Trees

Extra Trees(Extreemly Randomized Trees)는 앙상블 의사결정 트리 기법을 기반으로 한다. 랜덤 포레스트와 유사하게 랜덤하게 feature들을 선택하여 노드 분할에 활용한다. 그러나 각 노드에서 최적의 분할을 찾는 대신 무작위로 분할을 진행한 후 그중 최적을 선택한다.

[하이퍼파라미터]

n_estimators: 트리의 개수를 지정한다. 높은 값을 선택하면 모델의 복잡도가 증가한다.
max_features: 각 노드에서 무작위로 선택되는 feature의 개수를 제어한다. 값이 작을수록 모델의 다양성을 증가시키고 과적합을 줄인다.
max_depth: 의사결정 트리의 최대 깊이다. 높을수록 모델의 복잡도가 증가한다

- Random Forest

앙상블 기법 중 하나로 의사결정 트리를 기반으로 한다. 여러 개의 트리를 조합하여 모델을 형성하고 개별 트리의 약점을 보완하고 안정적인 예측 수행이 가능해진다. 높은 예측 성능을 발휘하며 변수의 스케일 조정이나 정규화 없이도 고차원 데이터에 잘 적용이 가능하다. 또한 마찬가지로 변수의 중요도를 평가하여 어떤 특성이 예측에 중요한 역할을 하는지 확인할 수 있어 해석력이 높다.

[하이퍼파라미터]

n_estimators: 생성되는 의사결정 트리의 개수
max_features: 각 노드에서 무작위로 선택되는 특성의 개수
max_depth: 의사결정 트리의 최대 깊이 제한

- XGBoost(Extreme Gradient Boosting)

Gradient Boosting 알고리즘의 확장 형태이다. L1(Lasso) 및 L2(Ridge) 규제를 통해 모델의 복잡도를 제어하여 과적합을 방지하고 기존의 Gradient Boosting보다 더 빠른 학습과 나은 예측 성능을 제공한다. 마찬가지로 변수의 중요도를 평가하여 모델 해석력을 높이며, 결측치 처리를 자동으로 지원해서 누락된 데이터를 적절하게 다루고 모델 예측 능력을 향상시킨다.

[하이퍼파라미터]

n_estimators: 트리의 개수 지정. 클수록 모델 복잡도가 증가한다.
learning_rate: 각 weak learner의 기여도 조절 파라미터.
max_depth: 의사결정 tree의 최대 깊이 제한

***앙상블**: 여러 개의 분류기를 생성하고 그 예측을 결합함으로써 더 정확한 예측을 도출하는 기법이다. 강력한 하나의 모델을 사용하는 대신 성능이 상대적으로 떨어지더라도 모델을 여러 개 조합하여 서로 보완하고 더 좋은 예측력을 갖도록 한다. 단일 모델을 활용하는 것보다 각 모델의 장점을 반영하면서 보다 안정화된 모델을 만들 수 있게 된다.

- K-fold Cross Validation

머신러닝 모델의 성능을 평가하고 일반화 능력을 검증하기 위한 효과적인 기법 중 하나다. 데이터를 k개의 부분집합으로 나누고 k번의 실험을 통해 모델을 학습 및 평가하는 방법이다. 각 실험별로 한 개의 부분집합을 검증 데이터로 사용하고 나머지 부분집합을 학습 데이터로 사용했다. 여러번의 실험 결과를 통해 모델의 성능을 평가하고 최종적인 성능 지표를 얻을 수 있게 된다. 데이터를 여러 부분으로 나워 실험을 반복적으로 진행한다는 점에서 충분한 검증력을 제공한다.

-Isolation Forest

이상치 탐지 및 제거를 위해 활용된 알고리즘이다. 작동원리는 먼저 랜덤 분할을 통해 무작위로 데이터 feature를 선택하고, 선택된 feature를 통해 데이터를 분할한다. 이때 분할 횟수를 눈여겨본다. 이상치 데이터의 경우 일반적인 데이터 포인트보다 훨씬 적은 분할 횟수를 가지는 경향이 있다. 따라서 트리의 깊이가 이상치는 일반 데이터에 비해 낮을 수 밖에 없다. 분할 횟수를 측정하여 Iso Forest는 각 데이터 포인트가 몇 번의 분할을 거쳐 노드에 도달하는지 이상치 점수를 계산한다. 해당 값이 낮을수록 이상치로 간주될 확률이 높아지고, 일정 임계값을 초과하는 데이터 포인트를 제거함으로써 이상치 식별을 마친다. Isolation Forest는 특히 고차원 데이터셋에서 빠른 계산 속도로 효과를 발휘하는 장점이 있다.

(2) 군집화 과정에서 활용한 분석 기술

-K-Means

주어진 데이터를 k개의 군집으로 묶는 알고리즘이다. 각 데이터 간 서로에 대한 유사성을 기반으로 k개의 군집을 찾는다. 이때, 사전에 군집의 개수를 유저가 정의해야 한다. 한번 분리된 개체도 반복적으로 시행하는 과정에서 재분류된다. 또 관측치들 사이 유사성(거리)를 이용하여 거리 차이의 분산을 최소화하는 방식으로 동작한다. 군집 결과는 서비스 차원에서 고객들에게 페르소나를 부여하기 위해 사용되었다.

(4) 변수 해석시 활용한 분석 기술

-> 영향력 있는 변수를 사람이 임의적으로 개입하여 주관적으로 판단하는 것이 아니라, 머신러닝 알고리즘과 사후 분석을 통해 모델이 스스로 판단하도록 방식을 채택하였다.

(5-1) Feature Importance

독립변수 중에서 종속변수인 사고율을 예측할 때 유의미한 정보가 가장 많이 포함된 변수들을 순차적으로 파악할 수 있는 시각화 기법이다. Tree 기반 모델을 활용하면 함께 결과로 제시된다. 모델에 의해 계산된 특성 중요도를 나타내는 수치를 막대 그래프로 알아보기 쉽게 표시한다. 이때, 각 특성에 대한 중요도는 해당 특성에 예측 모델의 결과에 얼마나 영향을 미치는지를 나타낸다. 즉, 예측에 중요한 특성을 식별하는데 큰 도움이 된다. 또 반대로 예측에 큰 영향을 주지 않는 불필요한 변수들도 확인하여 개선 작업을 수행할 수 있다.
핵심은 모델링 과정에서 산출되는 변수 중요도를 해석할 수 있다는 점이다. 이는 종속 변수를 예측할 때 유의미한 정보를 가지고 있는 변수를 파악하는 과정이다. 파생 변수들의 효력을 검증하고, 새로운 파생변수들 간의 상호작용도 파악할 수 있다. 변수 간 중요도를 내림차순으로 출력하여 종속변수인 사고율 예측에 유의미한 정보가 담긴 변수를 도출해낼 수 있다.

(5-2) Partial Dependency Plot

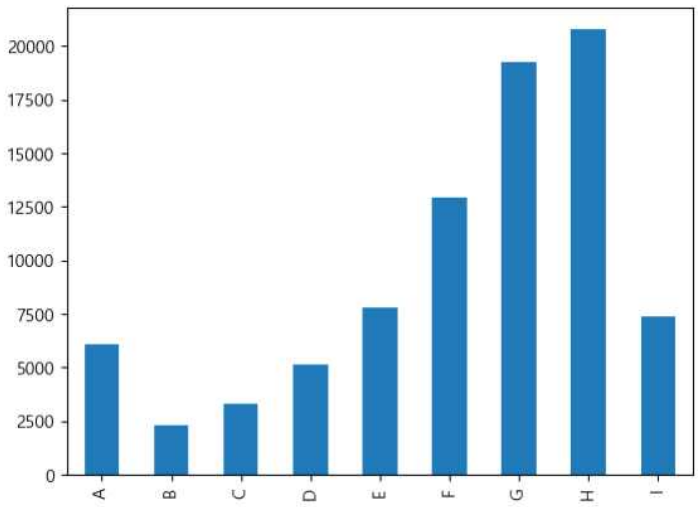
모델의 예측값에 대해 하나의 feature의 한계효과를 나타내는 기법으로 타깃과 특성 간의 관계가 선형인지 단조함수인지 파악하게 해준다. 모델에서 특정 feature가 최종 사고율 예측에 어떤 수치적인 영향(Marginal Contribution)을 끼쳤는지 시각적으로 확인할 수 있다. Feature Importance와의 차이점은 특정 feature 값이 변할 때 모델의 예측 결과 또한 어떻게 변화하는지를 보여준다는 것이다. 경우에 따라 특정 feature와 예측 결과 간의 비선형 관계를 파악할 수 있도록 도와주기도 한다.
PDP 그래프를 출력하면 사고율에 영향을 끼치는 변수의 영향력이 제시된다. 하나의 독립변수가 최종 사고율 예측에 평균적으로 어떤 영향을 끼쳤는지 수치적으로 드러난다. 또한 독립변수 하나의 개별기여도도 확인 가능하다.

Ⅲ. 데이터분석 및 모델링 과정

데이터 분석 및 모델링 과정에 대한 전반적인 과정에 대해 설명하겠다.

(1) 데이터 수집 및 EDA

- 변수들 설명, 개수
- 소득(INCOME) 변수의 시각화



- accident_flag의 변수 분포도 -> 불균형 데이터셋임을 확인

생명보험(1)
False 16360
True 17
장기보험(2)
False 14035
True 379
자동차보험(3)
False 52970
True 1240

- 기술 통계
- 상관관계 분석모델링에 앞서 어떤 부분이 전처리되어야 하고 해결해야하는지 판단 (데이터 불균형 처리 및 파생변수 생성으로 새로운 관계성을 찾아야 함)

(2) 데이터 전처리

(2-1) Feature Engineering (파생변수 생성)

원본 데이터셋에 더 다양한 관점에서 학습이 이뤄질수 있도록 여러 파생변수를 새롭게 형성하였다. 다음은 형성된 파생변수들이다:

- accident_flag: 가장 중요한 파생변수다. 정의는 “사고 유무”인 categorical 변수이며, 1은 사고가 단 한번이라도 발생한 경우 그리고 0은 그 반대로 사고가 한번도 발생하지 않은 경우를 뜻한다. LOSS(손해액)>0 또는 ACCD_CUNT(사고건수)>0 인 경우로 산정한다. 본래 종속변수인 사고율을 대체하기 위한 지표다.

다음 변수들은 집 주소와 출근지 간의 분리된 시도와 시군구 주소를 합쳤다.

- HOM_ADDR: HOM_MGPO와 HOM_SSG를 합쳐 형성하였다

- OFFI_ADDR: OFFI_MGPO와 OFFI_SSG를 합쳐 형성하였다

다음 변수들은 집 주소와 출근지가 같은지 여부를 확인하였다

- HOM_OFFI_MGPO_SAME: HOM_MGPO == OFFI_MGPO 여부 확인

- HOM_OFFI_SSG_SAME: HOM_ADDR == OFFI_ADDR 여부 확인

시간 정보도 포함하여 어느 시간대에 고객의 정량적인 지출이 높은지 확인 가능하다.

- 매출금액합계_총

- 매출건수합계_총

- 매출건당금액_총: 매출금액합계_총/매출건수합계_총

- 매출금액합계_오전

- 매출금액합계_오후

- 매출건당금액_오전: 매출금액합계_오전/매출건수합계_오전

- 매출건당금액_오후: 매출금액합계_오후/매출건수합계_오후

(2-2) Feature Elimination

Feature Engineering에 활용된 변수들을 파생변수들이 생성된 후에도 모델링에 계속 남겨둔다면 과적합 및 차원의 저주 문제가 발생한다. 따라서 사용하거나, 예측에 도움이 별로 되지 않을 것으로 판단되는 변수들은 제거하였다. 다음은 drop한 변수명들이다:
'PREM', 'LOSS', 'ACCD_CUNT', 'INSR_TYPE', 'accident_flag', 'HOM_MGPO', 'HOM_SGG', 'OFFI_MGPO', 'OFFI_SGG'

(2-3) Categorical Variables (변수 인코딩)

범주형 변수로 판단되는 변수들은 변수 type을 category로 변환하였다. 다음이 categorical 칼럼명들이다:

'SEX', 'JOB', 'INCOME', 'HOM_ADDR', 'OFFI_ADDR', 'EST_LFSTG', 'HOM_OFFI_MGPO_SAME', 'HOM_OFFI_SGG_SAME'

(2-4) 데이터셋 분할

보험 종류(INSR_TYPE)에 따라 데이터셋을 각각 생명보험(1), 장기보험(2), 자동차보험(3)으로 분류하였다. 또한 변수들도 구분하여 insr_col과 card_col으로 각각 보험개발원 데이터와 삼성카드 데이터도 분류하였다.

(2-5) 불균형 데이터 처리

데이터 불균형은 특정 클래스의 데이터가 다른 클래스에 비해 월등히 많거나 적은 경우를 뜻한다. 불균형을 유지한채로 학습을 진행하면 모델은 주로 클래스가 많은 부분에 초점을 맞춰 비대칭적인 학습을 수행한다. 결국 소수 클래스의 패턴을 제대로 학습하지 못하고 모델 성능이 저하되는 측면이 있다.

실제 데이터를 살펴보면 "사고건수" 변수가 0인 것이 꽤 많다. 통상적으로 사고에 해당되는 데이터 수는 그렇지 않은 정상 건수(사고율이 0인 데이터)에 비해 현저히 적기 때문이다. 때문에 이를 처리하지 않고 모델링을 진행할 시, 예측 정확도를 높일수 있는 최적의 모델을 찾는데 방해가 된다. 그리고 accuracy와 같은 데이터가 다소 균형을 이룬 상태에서 활용되는 평가 metric은 쓰지 못하게 된다.

데이터 불균형을 처리하기 위한 대표적인 기법으로 샘플링(Sampling)이 있고, 대표적인 기법으로 주류 데이터를 조절하여 소수 데이터 수와 동일하게 만드는 Undersampling 기법, 그리고 반대로 소수 데이터의 샘플을 복제하여 전체 데이터 간의 균형을 맞추는 SMOTE 등의 oversampling 기법도 존재한다. 본 데이터셋에서는 undersampling 방식을 활용하여 불균형 처리를 최종적으로 마무리하였다.

[3] 예측 모델링

예측 모델링의 핵심 목표는 2가지다

1) 마이데이터의 카드 거래 내역을 추가하여 더 정교하고 정확한 예측 성능의 모델을 개발하는 것
2) 제작한 변수들의 영향력과 중요도에 대해 해석 가능한 머신러닝 모델 만들기
정리하면, 성능도 어느정도 보장되면서 설명가능성이 높은 모델을 제작하고자 한다. 또한 보험의 종류들도 고려하여 개별 보험 종류별(1:생명, 2:장기, 3:자동차)로 모델링을 시도하였

다.

모델링의 순서는 다음과 같이 여러 후보 모델 제안 -> 최적의 모델 선택 -> 해당 모델로 학습 진행된다.

(3-1) 후보 모델 선택

분석 목적 중 하나인 “설명가능성/결과 해석력”을 고려하여 트리 기반 모델을 모델 후보들로 선정하였다. 구체적으로 총 5가지 모델 종류를 고려하였다: Gradient Boost, ExtraTree, RandomForest, XGB, Decision Tree

아무런 튜닝 과정을 거치지 않은채 먼저 5개의 Baseline Model을 바탕으로 A) 보험개발원 데이터셋과 A+B) 보험개발원+카드 데이터셋 각각 2개에 5개의 모델을 모두 적용하여 기본 성능을 평가했다. 이때, cross-validation 기법(cv=5)을 통해 모델들의 평균적인 성능을 구하였다. 성능 결과는 다음과 같다:

<장기보험(2)>

A) 보험개발원 데이터셋
gb 0.7553231486303432
et 0.7008422413594281
rf 0.723147298211134
xgb 0.7421124718461958
dt 0.7051878433258587

A+B) 보험개발원 + 카드 데이터셋
gb 0.7290176706191434
et 0.5975622735228647
rf 0.6324900134559981
xgb 0.7220459637920142
dt 0.6442476679933435

<자동차보험(3)>

A) 보험개발원 데이터셋
gb 0.7420625001257575
et 0.6492143335339594
rf 0.6789382102641598
xgb 0.7103747724651769
dt 0.6289090595082789

A+B) 보험개발원 + 카드 데이터
gb 0.7420625001257575
et 0.6530415334676081
rf 0.6817501840439368
xgb 0.7103747724651769
dt 0.6315569587303829

최종 학습 모델을 선택하는데 있어 다음과 같은 여러 기준을 고려하였다:

1) 성능 2) 해석 가능성 3) 패키지 사용 편의성 4) 학습 소요 시간

해당 기준들을 종합하여 최종 모델로 XGBoost Classifier를 선정하였다. 성능에 있어서는 Gradient Boost가 가장 뛰어나지만 XGB도 못지 않게 성능이 좋고, 무엇보다도 "해석 가능성"에 있어서 Feature Importance Plot을 활용할 수 있어 직관적으로 결과를 해석할 수 있다.

(3-2) 모델 학습 및 평가

모델 학습 과정의 경우 크게 1) 생명보험과 2), 3) 장기&자동차 보험으로 나눠 진행되었다. 데이터 leakage를 방지하기 위해 Train/Test 데이터로 나눠서 진행이 되었다.

[1] 생명보험

사고유무 변수를 예측하기 위해 수행해야 하는 task는 binary classification이다. 이진 분류 문제에 있어 가장 널리 활용되는 로지스틱 회귀분석 기법을 활용하였다. 총 31개의 feature를 바탕으로 로지스틱 모델을 형성하였고 TOP 6개의 important feature를 도출하였다. 최종 예측 성능으로 약 0.8710의 f1-score가 나왔다. 이때 사고율을 대체하기 위한 값을 찾기 위해 사고유무 변수를 classification으로 모델링하여 결과로 산출된 확률값을 사고율의 대체변수로 결정하였다.

[2] 장기보험 & 자동차보험

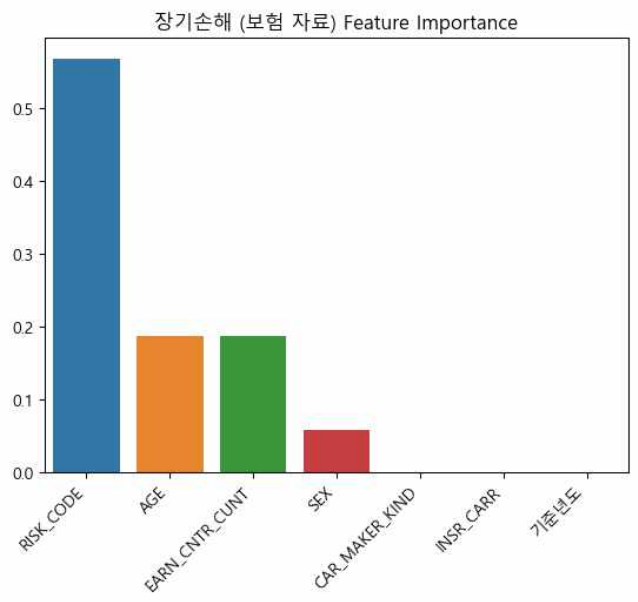
두 보험의 경우 카드 데이터를 추가적으로 활용했다. 보험개발원(A) 데이터로만 학습한 모델과 보험+카드 데이터(A+B)로 학습한 모델들간 비교가 진행되었다. 따라서 총 4가지 모델(장기: model_2_a vs model_2_ab / 자동차: model_3_a vs model_3_ab)이 완성되었다. 마찬가지로 각 모델별로 feature importance 분석을 진행하였다.

해당 모델들에서는 feature 제거 및 이상치 제거의 효과를 보기 위해 XGBClassifier를 생성하였다. SelectFromModel을 통해 pretrained된 XGB모델을 불러오고, IsolationForest를 기반으

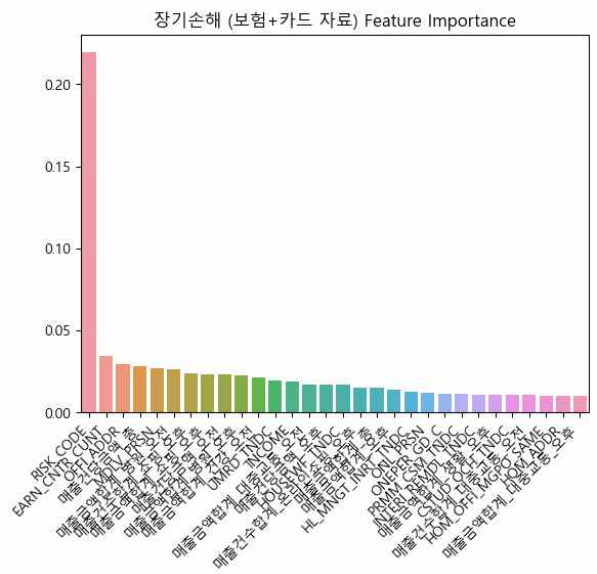
로 이상치 제거를 실시하였다. 사전 작업이 완료된후 모델 학습을 진행한 결과는 다음과 같다:

model_2_a: 0.7368 / model_2_ab: 0.7158
model_3_a: 0.7951 / model_3_ab: 0.6806

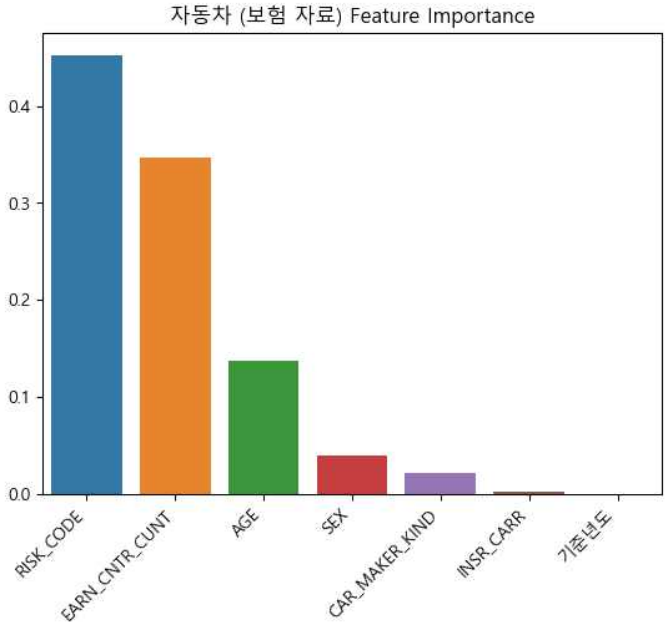
<중요 Feature>
model_2_a



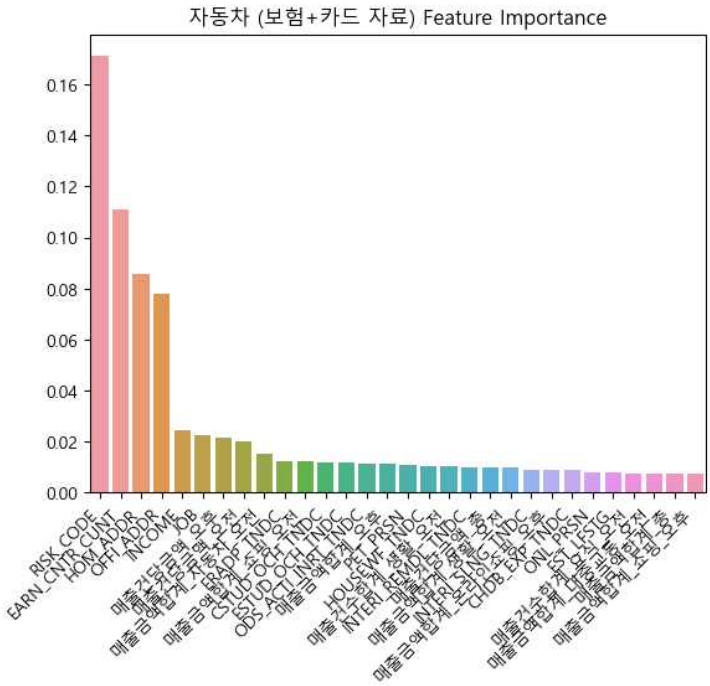
model_2_ab



model_3_a



model_3_ab



평가 지표: F1_Score

머신러닝에서 분류 모델의 성능을 평가하기 위해 사용된 지표다. 정밀도(Precision)와 재현율(Recall)의 재현율의 조화 평균값으로 계산된다. 해당 지표를 accuracy 대신 사용한 이유는 데이터셋의 클래스 불균형 및 변수들의 불균형한 분포를 감안한 결과다.

F1 Score = 2*Precision*Recall/(Precision+Recall)
Precision = TruePositive/(TruePositive+FalsePositive)
Recall = TruePositive/(TruePositive+FalseNegative)

(4)군집화

모델링을 완료한 후, 카드 데이터를 추가적으로 이용해 예측한 사고율이 보험 데이터만을 이용해 예측한 사고율보다 성능이 좋다는 결과가 도출되었다면, 각 고객을 예측사고율과 실제사고율을 바탕으로 군집화하게 된다. 이때 예측 사고율은 보험데이터만을 사용한 사고율이 되고, 실제사고율의 경우 더 높은 성능이 도출된 카드데이터로 이용한 사고율이다. 각 사고율을 x,y축에 놓고 데이터의 분포를 확인하며 고객의 군집을 총 4개로 나눌 것이다. 이때, 군집의 개수를 직접 지정할 수 있는 유클리드 거리 기반의 군집화 방식인 K-Means를 사용할 것이다. 이를 통해 각 군집이 효과적으로 분포되어있나 확인한 후, 각 집단의 특징에 대해 알아보며 고객 유형별 전략을 고민해볼 것이다.

Ⅳ. 데이터분석 및 모델링 방법론

데이터 분석 및 모델링 방법론을 고려하는데 있어 다음 2가지를 중시하였다:

1. **모든 분석 및 모델링 과정에서의 파라미터, 결과값 등은 데이터 및 알고리즘이 알아서 스스로 판단하게끔 하였다. 사람의 개입은 최소화하였다.**
2. **단순 AI 모델의 성능 높이기에 초점을 두지 않았다. 모델의 핵심 목적을 고려하여 예측 정확도(성능)과 모델 설명력(해석력)을 함께 높이고자 했다.**

(1) 데이터 수집 및 EDA

(1-1) 데이터 출처

주어진 데이터 외에 추가적인 외부 데이터는 활용하지 않았다. 우선 이미 기본적으로 주어진 데이터셋이 크고 활용 가능한 변수가 충분함을 고려하여, 외부 데이터를 추가할 경우 데이터셋의 차원이 너무 커질 것이라 판단했다. 이미 기존 변수의 활용만으로도 충분히 해석 가능하고 성능이 좋은 모델을 만들 수 있을거라고 기대했다.
두 번째로 본 모델을 통해 검증하고자하는 프로젝트의 목적을 고려하였다. 모델을 통해 마이데이터의 카드 거래 내역이 사고율 예측에 도움 되는지 가설을 검정해야 한다. 더 엄밀하게 말하면, 기존의 사고율 예측 방식(A 데이터만 활용) vs 새 모델의 사고율 예측(A+B 데이터셋 활용)를 비교해야 한다. 때문에 여기서 새로운 추가 데이터를 활용한다면 해당 방식으로는 우리가 검증하고 싶었던 가설의 결과를 확인하기 어렵다.

(1-2) EDA

데이터 시각화 및 기술 통계 등 EDA를 진행한 이유는 추후 전처리 및 모델링 과정에 있어서의 탐색 후 인사이트를 얻기 위해서다.
1. 다각적인 시각에서 고객의 행동패턴을 파악하기 위함.
2. 모델 훈련시 데이터 비율을 조정하여 리샘플링을 통해 데이터 불균형을 해결하기 위함.
3. 파생 변수 생성에 있어 종속 변수와 상관관계가 비교적 높은 변수를 도출하기 위함.

(2) 데이터 전처리

(2-1) Feature Engineering (파생변수 생성)

Feature Engineering을 진행하는 이유는 기존 변수들의 활용만으로는 보다 더 정교한 사고율 예측을 수행하기 어렵다. 따라서 건강, 소득, 시간대 등 다양한 관점에서 보험 분야의 비즈니스 맥락에 맞게 새로운 파생변수를 형성하여 모델이 feature들을 다채롭게 학습할 수 있도록 취지가 있다. 해당 과정은 또한 도메인 지식을 요구하고 반영한다는 점에서 추후 모델링에 의한 결과 해석 과정에서도 큰 도움이 된다.

분석에 있어 가장 처음으로 맞닿아트린 문제가 주어진 데이터셋으로부터 종속변수인 사고율을 도출하는데 있어서였다. 기본적으로 데이터셋에 결측값이 많았고, 특히 자동차 보험 데이터의 경우 보험료인 PREM 값에 0이 많았다. 이는 기존 보험사들의 사고율 산정 공식인 "지급금액/보험료"에서 분모인 보험료가 0으로 취급되는 경우이기에 사고율을 궁극적으로 계산할 수가 없다. 따라서 기존 공식을 활용하지 않고 종속변수에 대한 feature engineering을 통해 새로운 대체지표인 accident_flag를 만든 것이다. "사고 유무"를 가진 해당 변수를 종속 변수로 예측에 활용하여, 유사한 지표를 통해 사고율을 간접적으로나마 확률적인 접근을 통해 분석을 시도하였다. 즉, 사고율을 새롭게 정의한 것이다.

Feature Engineering을 통해 누릴 수 있는 효과는 다음과 같다:

1. **차원 축소: 고차원 데이터에서 유용하지 않은 특성을 제거하거나 합치는 등의 처리를 통해 차원을 감소시킨다. 이로써 모델의 복잡성을 줄이고 계산 효율성을 향상시켜 차원의 저주(Curse of Dimensionality) 문제를 해결할 수 있다**
2. **과적합 방지: 특성 엔지니어링을 통해 데이터에 내제된 정보를 더 잘 추출할 수 있고, 궁극적으로 모델의 일반화 능력을 향상시켜 안정성을 높일 수 있다.**

추가적으로 차원 축소의 일환으로 대표적인 축소 기법인 PCA를 활용하는 것도 초기에 고려했으나, 모델 결과의 "해석 가능성"을 중시하는 관점에서 활용하지 않았다. PCA는 새로운 주성분이 원래 데이터의 선형 조합으로 구성되기 때문에 본래의 feature와의 관계가 어떻게 되는지 직접적으로 이해하기가 어렵다. 주성분 하나 하나가 feature와 일대일 대응이 되지 않기 때문이다. 즉, PCA를 활용하면 데이터 전반의 컨텍스트나 의미를 완벽하게 보존하기가 어려워 최종적으로 사용하지 않았다.

(2-2) Feature Elimination: 어떤 변수들을 제거하였는가?

과적합과 차원의 저주의 가능성을 우려하여 새로운 feature가 만들어지면, 사용했던 기본 변수들은 제거하는 것이 중요하다. 특히 예측 모델이 종속변수를 예측하는데 의미 있는 정보를 제공하지 못한 경우 변수를 마찬가지로 제거하였다. Feature Engineering과 마찬가지로 과적합을 방지라고 모델의 정확도가 올라가며, 추후 결과 변수 해석시도 용이하게끔 만든다.

(2-3) 결측치 처리: 변수별 중앙값으로 대체(Imputation with Median)

결측치를 처리하는 방식으로는 다양한 기법이 존재한다. 최우선적으로 결측치가 포함된 행들을 제거(Drop Missing Values)하는 방법론을 떠올랐으나 누락된 값들이 예상보다 너무 많이 포함되어 있었다. 제거를 하면 데이터셋의 행의 개수가 충분하지 않을 것으로 판단하여 데이터 손실을 우려해 제거는 진행하지 않았다.

회귀분석이나 KNN(K-Nearest Neighbors Imputation)과 같은 인접값들로 결측치를 보간하는

방법도 시도를 해보았으나, 이 역시도 이웃 데이터들에 결측치가 많이 포함되어 있어 진행하는데 제약이 있었다. 무엇보다도 다른 변수들 간의 관계가 아직 파악되지 않은 상태에서 해당 방법론들로 결측치를 대체하기엔 힘들어 보였다.

최종적으로 특정 값으로 결측치를 대체하는 방법을 택하였다. "평균(average)"로 대체를 첨에 생각했으나 연속형 변수들의 상당수가 정규분포가 아닌 skewed된 편향 분포임을 확인했다. 분포가 왜곡된 상태에서 평균으로 대체를 진행하면 변수들의 고유 특성을 제대로 반영하지 못하기에, 분포의 왜곡에 다소 강건한 "중앙값(median)"으로 최종 결측치들을 대체하였다. 해당 방법은 데이터의 분포를 어느 정도 고려하여 크게 왜곡하지 않으면서 결측치들ㅇㄹ 채워 넣을 수 있다는 장점이 존재한다.

(2-4) 이상치 처리: 진행 X

이상치 처리의 경우 따로 진행하지 않았다. 특정 데이터들이 이상치로 판별하기에는 불균형이 너무 심하고 대다수의 변수들이 skewed되어 있었기 때문이다. 더군다나 "기존 사고율과 새로운 모델이 예측한 사고율 간의 괴리"를 확인하고 분석하고자 하는 본 모델의 특성상, 분석 목적을 고려하여 이상치를 따로 처리할 필요는 없다고 판단하였다. 그리고 추후 모델링 과정에 있어 트리 기반 모델을 활용한 점을 고려하여, 트리의 branching을 통해 이상치의 영향력을 분산시켜 최소화할 수 있기에 이상치 처리가 별도로 필요하지 않았다.

(2-5) Scaling: 따로 진행 X

추후 모델링 과정에서 트리 기반 모델을 활용한 점을 고려하여 변수를 표준화시키는 스케일링 작업도 마찬가지로 따로 진행하지 않았다. 일반적으로 트리 기반 모델은 대부분의 데이터 스케일에 다른 머신러닝 알고리즘 대비 상대적으로 덜 민감한 특성을 지니고 있기에 따로 스케일링을 진행하지 않아도 분석에 지장이 없다. 이는 트리 기반 모델이 분할 기분을 찾는데 있어 feature의 절대적인 크기보다는 순서나 상대적 크기에 더 의존하기 때문이다.

(2-6) 불균형 처리: Undersampling

본 데이터셋이 대규몽 점을 고려하여 샘플 수를 오히려 늘리는 것보다는 줄여나가는 Undersampling 방법을 활용했습니다. 이를 통해 학습 비용 및 계산량을 현저히 감소시켰다. SMOTE를 비롯한 Oversampling 기법은 새로운 데이터를 지나치게 많이 생성하게 되어 원본 데이터가 너무 많이 복제되고 차원이 커지는 문제가 발생할 것으로 우려하여 활용하지 않기로 판단했다.

(3) 예측 모델링

(3-1) 사용한 트리 기반 후보 모델

트리 기반 모델을 사용한 이유는 다음과 같다: 보험료를 산정하고 고객에게 이를 설명할

때, 사고율이 어떻게 산정되었고 사고율에 영향을 끼치는 변수의 효과를 정확히 설명하는 것이 중요하다. 따라서 성능이 좋더라도 딥러닝과 같이 모델링 결과 도출 과정이 black box 인 모형들은 고객들을 설득하기에는 설명 가능성이 부족하다는 단점이 있어 고려하지 않았다. Explainable AI의 관점에서 비즈니스적으로 함의가 있는 모델, 그리고 해당 결과를 도출하기 위해 구체적으로 어떤 과정을 수행해야 하는지 모델이 제시할 수 있어야 한다. 구체적으로 어떤 feature가 중요한 영향을 끼쳤는지, 예측한 사고율과 실제값이 차이가 난다면 그 이유가 어떤 요인에 의해 발생했는지, 3) 모델이 보험사의 수익성에 기인하는지 등을 중시하며, 설명가능함을 목표로 하는 모형들을 사용했다.

의사결정나무에 입각한 모델은 모델링 시 연속적으로 발송하는 의사결정 과정을 시각화하여 해당 결정이 언제, 어떻게 이루어지는지 성과를 한눈에 볼 수 있어서다. 즉, 모델의 계산 결과가 직접 나타나기 때문에 “해석 가능”하다는 뜻이다. 모델을 선정하는 단계에서 총 5가지 후보군의 트리기반 모델을 고려하였다.

(4)군집화

군집화를 활용하는 목적은 고객들의 행동 패턴을 기반으로 통계적 군집 모델을 사용해 고객들의 주된 특성을 파악하고 이를 바탕으로 맞춤형 서비스 메시지를 제안하기 위해서다. 사고율이 비슷한 사람들끼리 군집이 형성될텐데, 해당 고객들간 어떤 공통점과 차이점이 존재하며 해당 고객들에게 필요한 맞춤형 서비스는 무엇일지 고민하는 비즈니스 인사이트 도출을 위해서다.

K-means를 선정한 이유는 알고리즘이 단순하고 빠르게 수행되며 계층적 군집(DBSCAN)보다 많은 양의 자료를 다룰 수 있기 때문이다. 본 데이터셋의 크기를 고려하여 k-means가 더 효과적일 것으로 판단했다.

K 값의 경우, 본 프로젝트의 목적에 따라 사전 도메인 지식을 반영하여 k=4로 결정하였다. k의 개수가 4인 이유는 “기존 사고율 예측 모델 vs 새 사고율 예측 모델”의 관계를 고려하여 총 4개의 고객 페르소나를 산정해야하기 때문이다. 해당 군집화의 결과는 하단의 결과 목차에서 서술되어있다.

Ⅴ. 최종결과

1. 결과 요약

보험별로 각 해당 가설에 대해 검증 결과를 제시하려고 한다.

[생명보험]

앞서 데이터분석 과정에서 서술한 Recursive Feature Elimination 방식을 활용하여 해석에 쓰일 6가지 변수를 추출하였다. 그 결과는 하단과 같다.

1.EAR_CNTR_CUNT
2.HOM_OFFI_SGG_SAME
3.매출건당금액_총
4.매출건당금액_오후
5.UMRD_TNDC
6.INTERI_SLNG_TNDC

해당 Feature을 사용해 로지스틱 회귀 모형에 적합한 결과 분류 Accuracy(정확도)의 성능은 약 93%로 상당히 높은 수준임을 알 수 있다. 이때, 생성된 예측 모형의 회귀계수를 이용하여 사고율에 대한 각 변수의 영향력을 파악할 수 있었다. 하단의 표는 변수별 회귀계수의 값을 내림차순으로 정렬하였다.

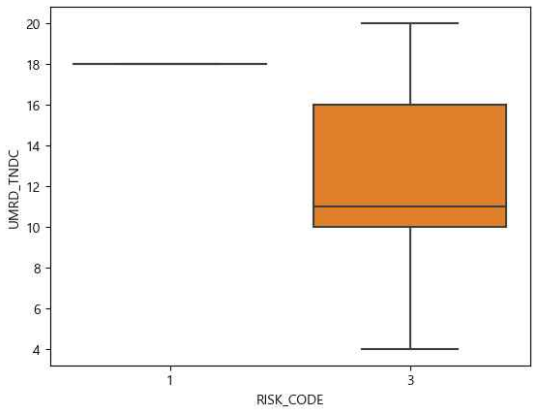
UMRD_TNDC	0.98652
매출건당금액_오후	0.77050
매출건당금액_총	0.56083
HOM_OFFI_SGG_SAME	-0.54064
INTERI_SLNG_TNDC	-0.54134
EARN_CNTR_CUNT	-1.47184

이를 바탕으로 UMRD_TNDC(미혼스코어), 오후 매출건당금액과 총 매출건당금액이 생명보험의 사고율에 양의 관계를 가지고 있음을, HOM_OFFI_SGG_SAME(거주지와 직장지의 광역시도와 시군구 모두 동일한 경우), INTERI_SLNG_TNDC(인테리어스타일링관심성향), EARN_CNTR_CUNT(보험 계약건수) 요인은 생명보험의 사고율에 음의 관계를 가지고 있음을 알 수 있다. 양의 관계를 가지는 3가지 변수의 경우, 미혼일수록, 카드결제 금액이 오후에 많을수록, 총 카드결제 금액이 많을수록 생명보험에서 사고가 더 많이 일어남을 확인할 수 있다. 역으로, 음의 관계를 가지는 3가지 변수의 경우, 거주지와 직장지의 광역시도와 시군구가 모두 동일한 경우, 인테리어 스타일링 관심성향이 높을수록, 보험 계약건수가 많을술

고 생명보험에서 사고가 더 적게 일어난다는 것을 확인할 수 있다.
하단의 그래프는 생명보험 내의 RISK_CODE별 위 6개의 변수 각각에 대한 분포를 나타낸 것이다. RISK_CODE가 2(재해사망)인 경우 해당하는 데이터가 없었기에 1(일반사망), 3(암발생)에 대한 해석을 진행하였다.

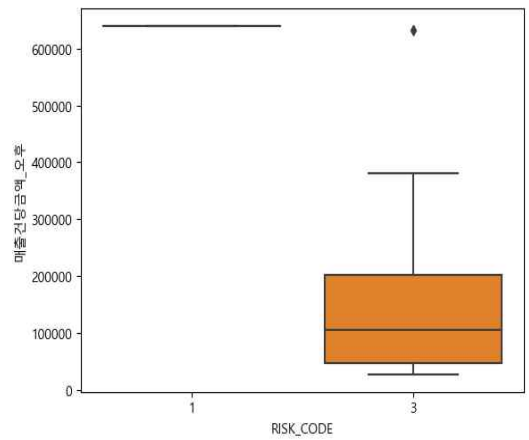
[미혼스코어]

암발생보다 일반사망에서 사고을 일으키는데 더 큰 영향을 가진다.



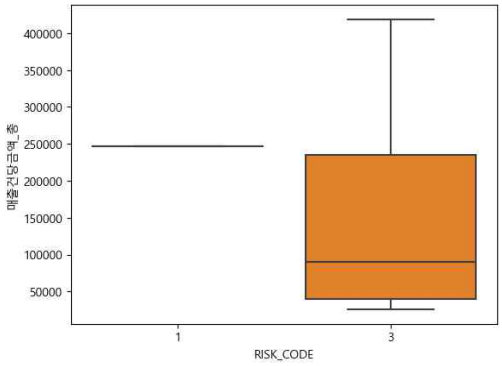
[오후 매출건당금액]의 경우

암발생보다 일반사망에서 사고을 일으키는데 더 큰 영향을 가진다.



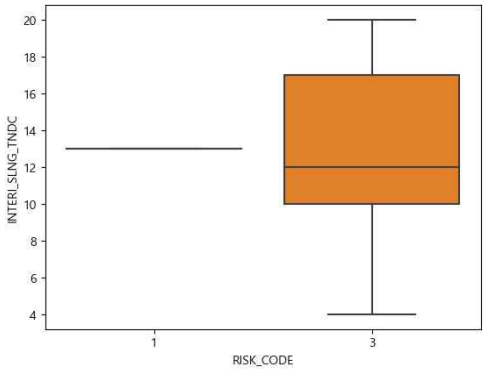
[총 매출건당금액]

암발생보다 일반사망에서 사고을 높게 유발하는데 더 큰 영향을 가진다.



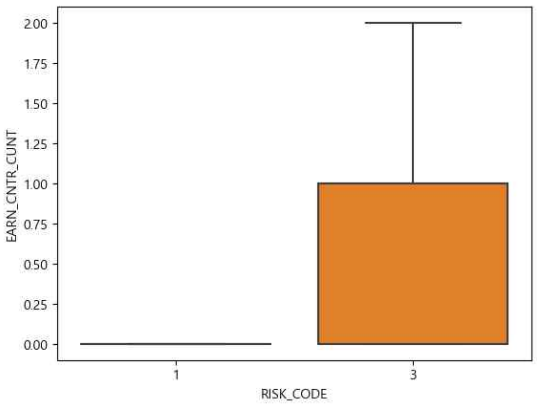
[인테리어 스타일링 관심성향]

암발생보다 일반사망에서 사고율을 적게 유발하는데 더 큰 영향을 미쳤다.



[보험 계약건수]

일반사망보다 암발생에서 사고율을 적게 유발하는데 더 큰 영향을 미쳤다.



[HOM_OFFI_SGG_SAME]
범주형 변수이므로 수치적인 평균으로 비교할 수 없어 보장내용별 해석에서 제외하였다.

이를 정리하자면, 암발생 보험과 비교했을 때, **일반사망에 대한 생명보험**의 경우 미혼스코어, 오후 매출건당금액, 총 매출건당금액이 높을수록 사고를 유발하는데 더 큰 영향을 가지고 인테리어 스타일링 관심 성향이 높을수록 더 작은 영향을 가진다.

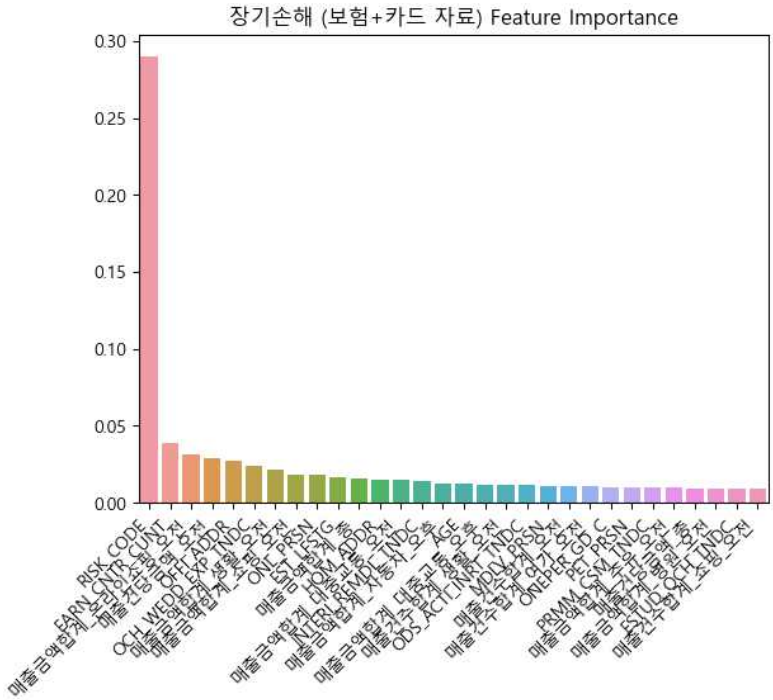
암발생에 대한 생명보험의 경우 오직 보험 계약건수의 변수에서만 일반사망 보험과 비교했을 때, 계약건수가 더 많을수록 사고율을 더 적게 유발하게 된다.

이처럼 사고유무에 영향을 끼치는 변수의 해석을 통해 가설2를 검증할 수 있다.

[장기손해보험]

보험 데이터만을 사용한 모델의 F1 Score는 0.69474, 카드 데이터를 함께 사용한 모델의 F1 SCORE은 0.76416으로, 결과적으로 카드데이터, 즉 마이데이터를 사고율 예측에 사용했을 때 사고율을 더 정확하게 예측하는 경향을 보였다. **이를 통해 가설1(마이데이터 데이터가 추가되면서 더 정확한 사고율 예측을 할 수 있다)이 입증됨을 확인할 수 있다.**

이때, Feature importance를 통해 사고율에 영향을 끼치는 주요변수를 파악해보았다.
그 결과는 하단의 그래와 같다.



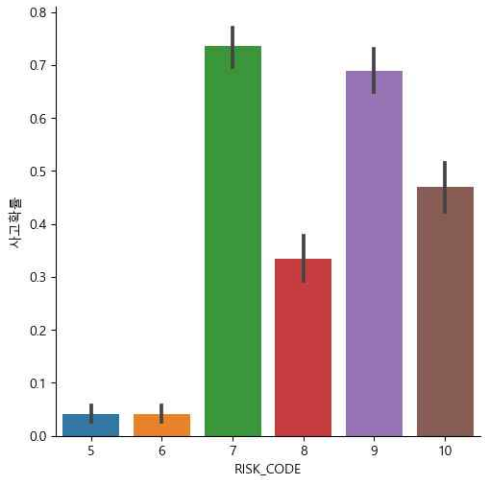
이때, 기울기가 급격히 하락하는 0.03을 기준으로 중요한 변수를 추출했을 때,

1. RISK_CODE 2. EARN_CNTR_CUNT 3. 매출금액합계_온라인쇼핑_오전

과 같은 주요한 3가지 변수를 파악할 수 있다.

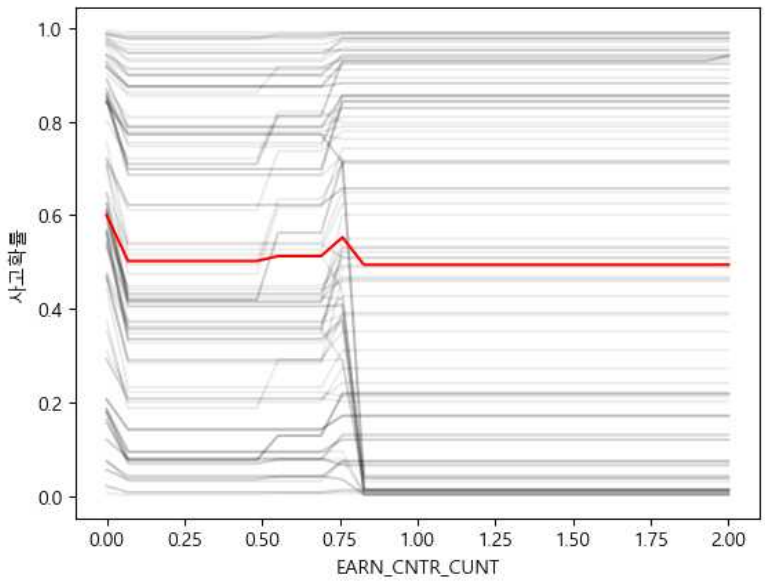
이를 토대로 장기손해보험에 대해 각 변수에 대한 해석을 진행해보기 위해 ICE curve와 PDP curve 기법을 적용하였다.

[RISK_CODE]
보장내용을 기준으로 질병수술>질병입원>상해입원>상해수술>급성심근경색진단=뇌졸증진단 순으로 사고가 날 확률이 높다는 것을 확인할 수 있다. 특히 하위 2개의 보장 내용인 급성심근경색진단과 뇌졸중진단의 경우 상위4개의 보장내용과 비교했을 때, 사고확률이 현저히 떨어짐을 확인할 수 있었다.



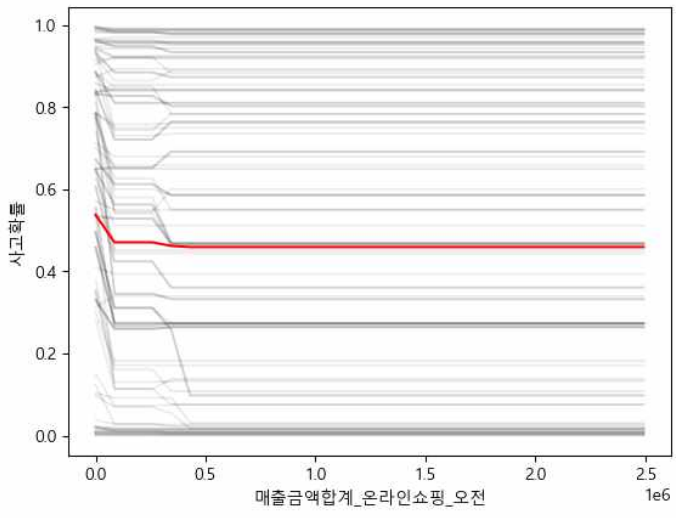
[EARN_CNTR_CUNT]

계약건수를 나타내는 해당변수의 경우 전체적인 그래프를 확인했을 때, 전체범위의 0~2에 따른 사고확률의 변화는 미비하다고 할 수 있다. 하지만 상대적으로 비교했을 때 계약건수가 1개 이상 있는 경우 나머지 경우보다 사고확률이 더 낮다는 것을 추론할 수 있다.



[매출금액합계_온라인쇼핑 오전]

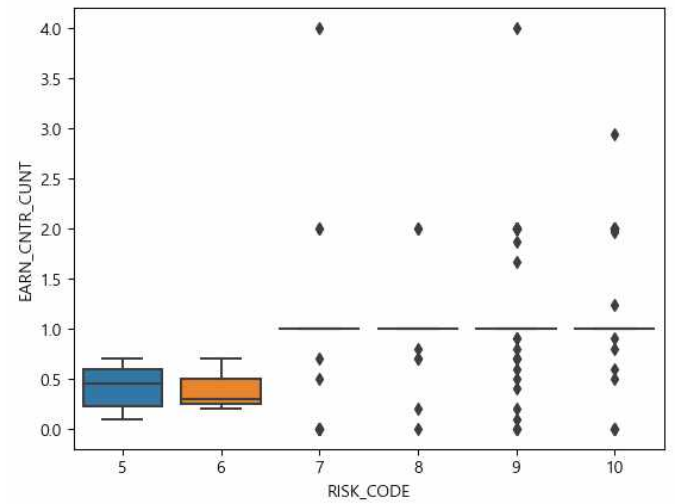
해당 변수는 오전 시간에 온라인쇼핑의 결제총금액을 나타낸다. 이 역시, 총금액에 따른 사고확률의 변화는 미비하지만, 평균값을 보여주는 빨간색 선을 기준으로 봤을 때, 총금액이 많아질수록 사고확률이 미비하게 낮아지고 있음을 확인할 수 있다.



다음으로는 장기손해보험의 RISK_CODE, 즉 보장내용별 변수의 영향력을 파악하고자 하였다. 이때 사용된 변수는 변수 중요도의 높은 영향을 보여주었던 하단의 2가지 변수를 사용했다.

1. EARN_CNTR_CUNT 2. 매출금액합계_온라인쇼핑_오전

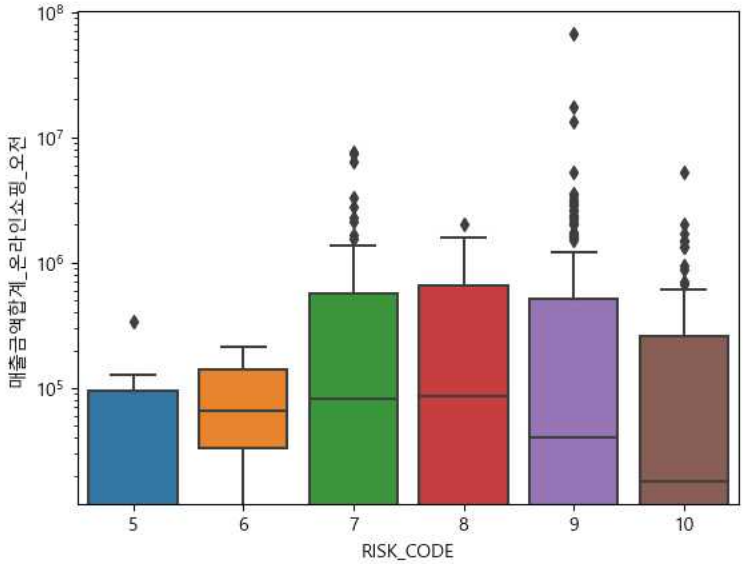
[EARN_CNTR_CUNT]



계약건수의 경우 10=9=7=8>5>6(해당 번호는 RISK_CODE) 순으로 해당값이 클수록 사고확률이 더 낮아짐을 확인할 수 있다. 즉 쉽게 말하자면, 상해입원, 질병입원, 질병수술, 상해수술이 뇌졸중진단, 급성심근경색진단 보장보다 계약건수가 더 많을 때 사고확률이 낮아질 확률이 더 높다는 것을 알 수 있다.

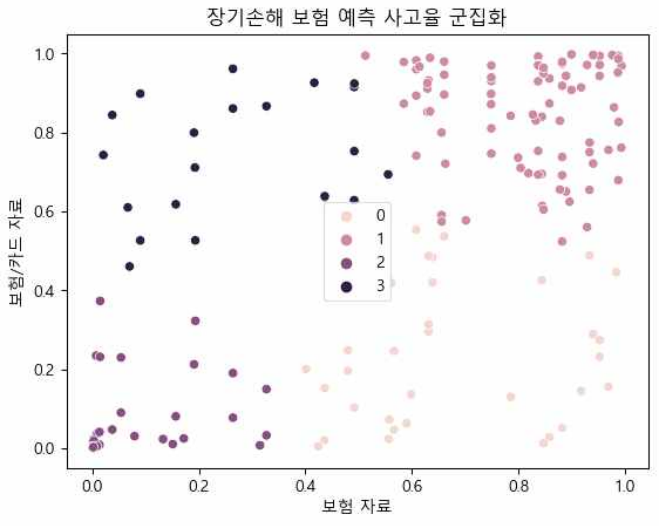
[매출금액합계_온라인쇼핑_오전]

오전 시간대 온라인쇼핑에서의 총결제금액 변수의 경우 8>7>6>9>10>5 순으로 해당값이 클수록 사고확률이 더 낮아짐을 확인할 수 있다. 즉, 상해수술, 질병수술, 질병입원, 급성심근경색진단에서 비슷한 중요도를 가지고 있고, 이는 뇌졸중진단과 상해입원 보장과 비교했을 때, 오전 시간대 온라인쇼핑에서의 총결제금액이 많을수록 사고확률이 더 줄어든다는 것을 확인할 수 있다.



해당 과정을 통해 가설2의 변수의 영향을 검증할 수 있었다.

추가적으로 장기손해보험의 경우 가설1에서 카드데이터를 추가사용한 모델이 보험 데이터만을 사용한 모델보다 사고율 예측에 더 효과적이라는 결론을 얻게 되었으므로 예상사고율과 실제사고율 간의 수치를 활용해 **군집화**하여 고객의 유형을 분류해볼 것이다. K-MEANS를 사용하여 군집화한 결과는 다음과 같다.



0, 1, 2, 3의 각각의 Label 값을 참고하여 고객의 유형을 4가지로 분류할 수 있다.

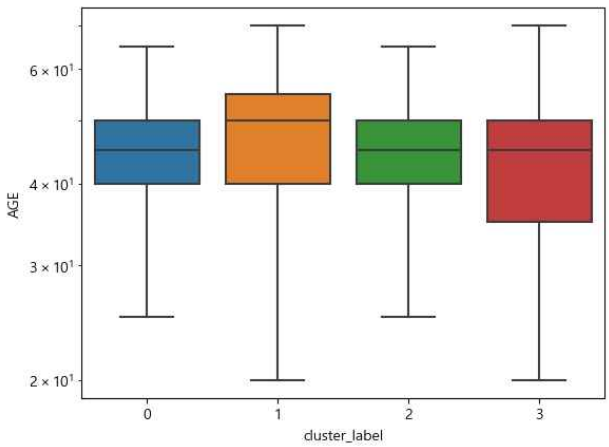
[1] 2: 예상 사고율이 낮고 실제 사고율도 낮은 고객
[2] 1: 예상 사고율이 높고 실제 사고율도 높은 고객
[3] 3: 예상 사고율이 낮고 실제 사고율이 높은 고객
[4] 0: 예상 사고율이 높고 실제 사고율이 낮은 고객

각 군집의 특성을 살펴보기 위해

1. AGE(나이) / 2. 총 매출건당금액 / 3. RISK_CODE / 4. INCOME(연소득)

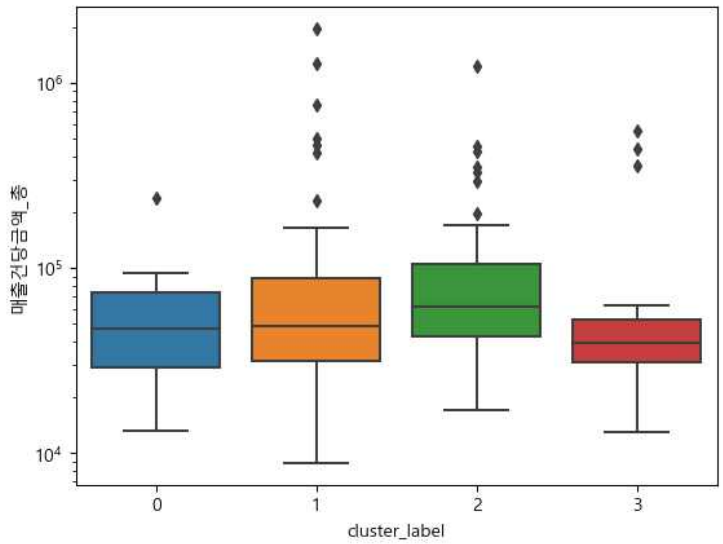
의 개인의 신상정보와 소비방식을 추론할 수 있는 대표적인 5가지 변수를 사용하였다.

[나이]
군집1의 경우 나머지 0,2,3과 비교했을 때 상대적으로 높은 나이를 가진 집단이다. 이때, 0,2,3 군집의 나이의 중위값은 비슷하다. 4개의 군집 모두에서 공통적으로 평균 나이대가 40~50대 사이인 것을 확인할 수 있다. 또한 1,3번 군집의 경우 나이대의 분포가 더욱 다양함을 알 수 있다.



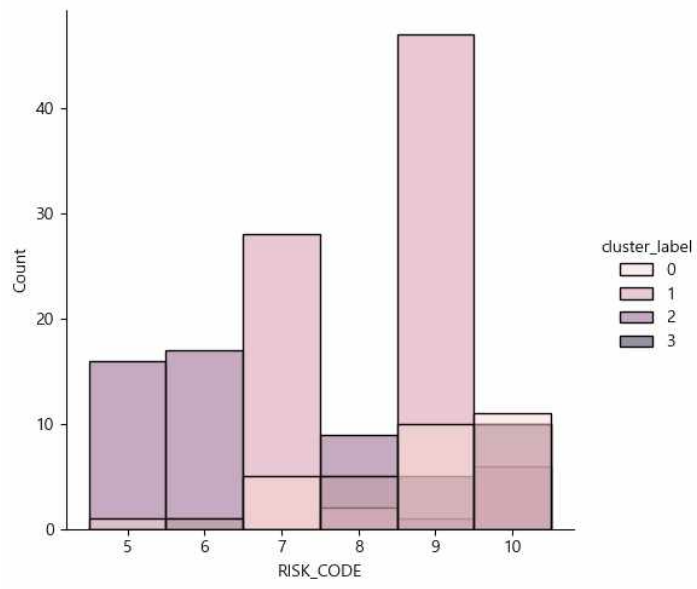
[총 매출건당금액]

총 매출건당금액의 경우 군집 2>1>0>3 순으로 높다. 하지만 그 절대적인 수치는 4개의 군집 모두 비슷하다. 특히 군집1과 군집2의 경우 이상치의 값이 상당히 많은 것을 확인할 수 있다.



[RISK_CODE]

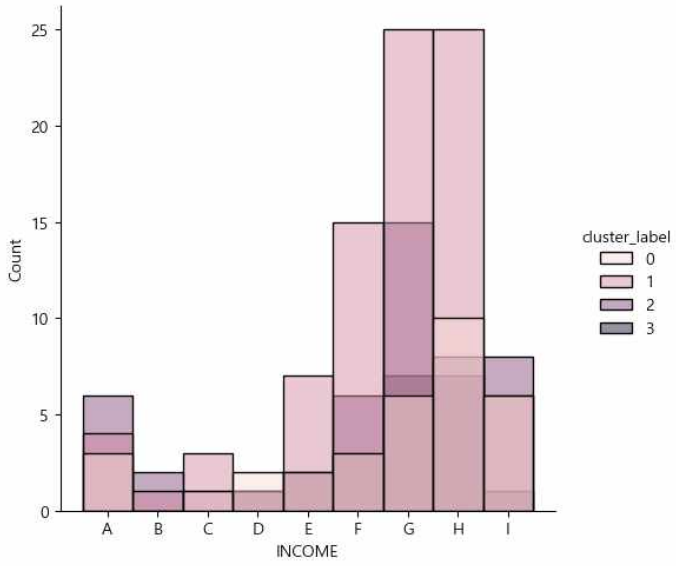
뇌졸중진단, 급성심근경색진단에 대한 보장 내용의 경우, 군집2가 대부분을 차지하고 있으며, 질병수술과, 질병입원의 경우 군집1의 사람들이 대부분 가입하고 있다. 또한 상해수술과 상해입원의 경우 군집2,3이 비슷하게 계약되어있다.



[연소득]

군집1,3의 경우 대부분 연소득이 6천만원 이하임을 알 수 있다.

전체적인 소득분포를 살펴보면 군집1>3>2>0 순으로 높은 소득을 가지고 있음을 알 수 있다.



이를 토대로 각 군집에 따른 고객의 특성을 종합해보면

군집0

평균나이대 40~50(그 중 최고령, 나이대 분포가 다양)

총 매출건당금액 3위

연소득 4위

군집1

평균나이대 40~50
총 매출건당금액 (2위, 이상치가 많음)
질병수술, 질병입원 보험 계약
대부분 연소득 (6천만원 이하, 1위)

군집2

평균나이대 40~50(나이대 분포가 다양)
총 매출건당금액 (1위, 이상치가 많음)
뇌졸중진단, 급성심근경색진단
연소득 3위

군집3

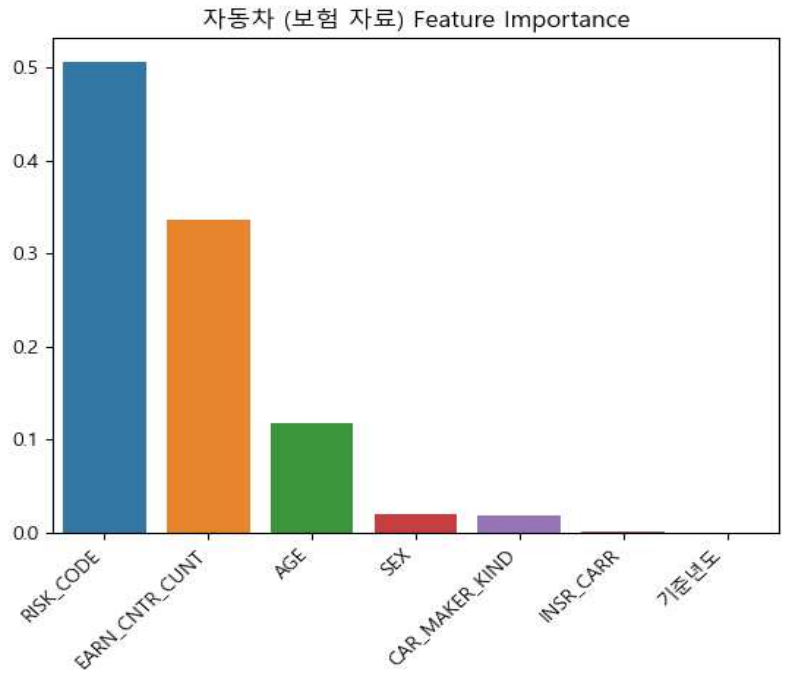
평균나이대 40~50
총 매출건당금액 4위
대부분 (연소득 6천만원 이하, 2위)

과 같다. EDA를 통해 다른 이외의 변수 역시 해석이 가능하다.
각 유형의 고객에 대한 CRM 전략의 경우 후에 '기대효과'에 서술할 예정이다.

[자동차보험]

보험 데이터만을 사용한 모델의 F1 Score는 0.78978, 카드 데이터를 함께 사용한 모델의 F1 SCORE은 0.65968로, 결과적으로 보험데이터만을 가지고 모형을 구축했을 때, 사고율을 더 정확하게 예측하는 경향을 보였다. **이를 통해 가설1(마이데이터 데이터가 추가되면서 더 정확한 사고율 예측을 할 수 있다)이 기각됨을 확인할 수 있다.** 해당 결론을 통해 추가적으로 자동차보험의 사고율과 카드 데이터 사이의 큰 연관성이 없다는 것을 추론할 수 있다.

그 후, Feature importance를 통해 사고율에 영향을 끼치는 주요변수를 파악해보았다.
그 결과는 하단의 그래프와 같다.



이때, 기울기가 급격히 하락하는 구간을 기준으로 중요한 변수를 추출했을 때,

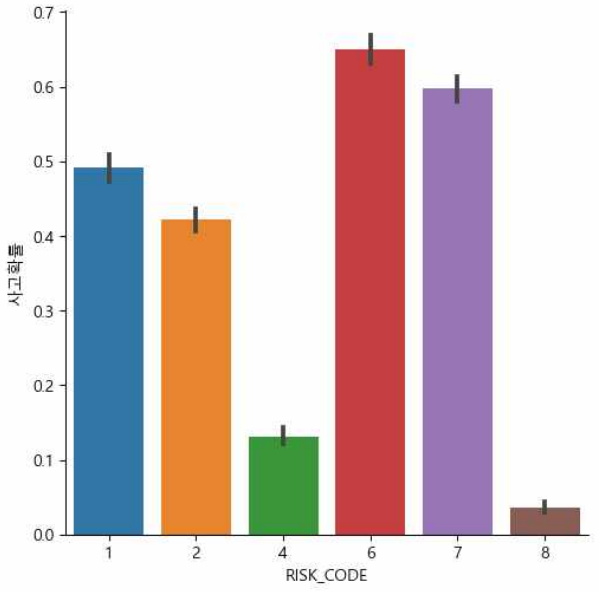
1. RISK_CODE 2. EARN_CNTR_CUNT 3. AGE

과 같은 주요한 3가지 변수를 파악할 수 있다.

이를 토대로 자동차 보험에 대해 각 변수에 대한 해석을 진행해보기 위해 ICE curve와 PDP curve 기법을 적용하였다.

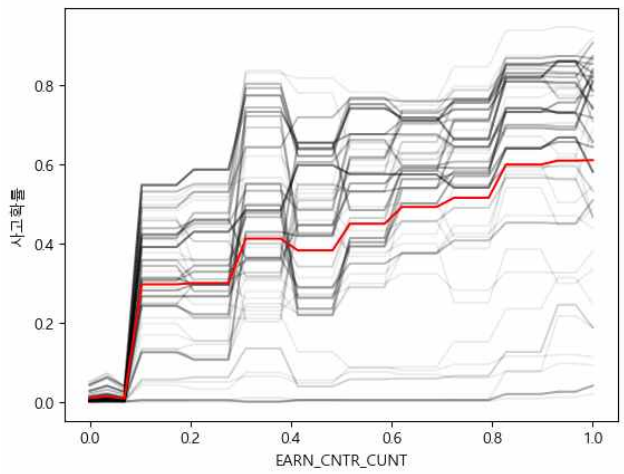
[RISK_CODE]

보장내용을 기준으로 **대물>자차>대인1>대인2>자손>무보험**순으로 사고가 날 확률이 높다는 것을 확인할 수 있다. 특히 하위 2개의 보장 내용인 자손과 자차의 경우 상위 4개의 보장내용과 비교했을 때, 사고확률이 현저히 떨어짐을 확인할 수 있었다. 대물 보장내용의 보험이 사고확률이 가장 많다는 것이 인상적이였다.



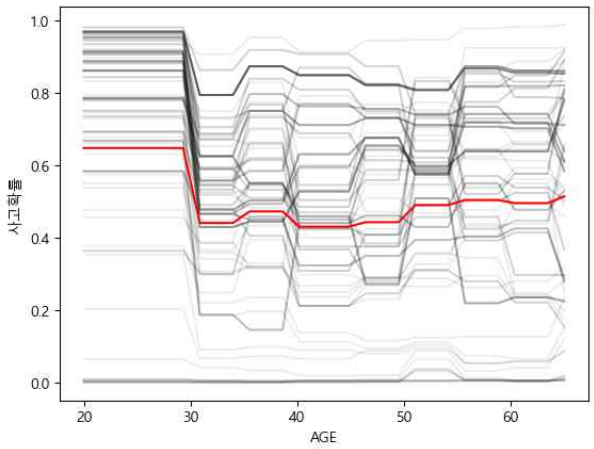
[EARN_CNTR_CUNT]

우상향하며 그 변동폭이 점점 줄어드는 그래프에서 알 수 있듯이, 계약건수가 증가할수록 사고확률 또한 증가함을 확인할 수 있다. 구체적으로는 계약건수가 0.7 이상일때 사고확률이 0.5 이상이 되어 사고가 날 확률이 높아짐을 알 수 있다.



[AGE]

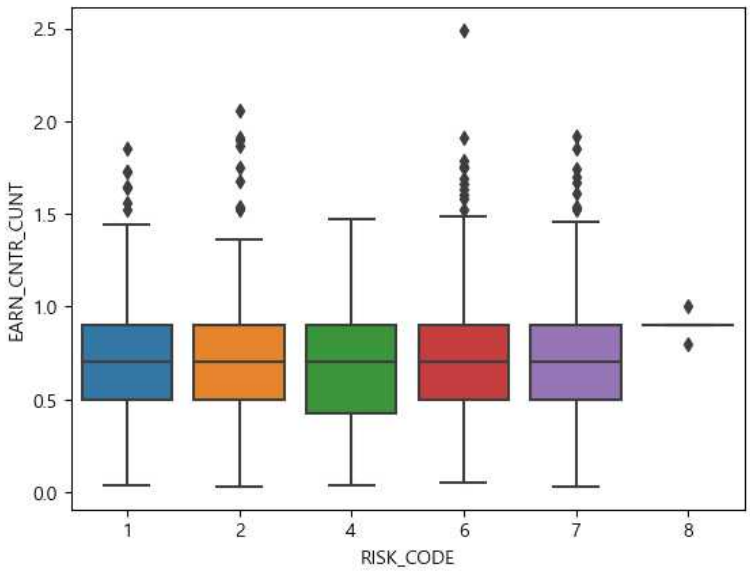
해당 그래프의 경우 20~30대에서 사고확률이 가장 높지만 그 이후부터 대폭 감소하여 50까지는 그 수준을 유지하다, 그 이후부터는 다시 사고확률이 높아지고 있다. 나이가 어린 초보운전자와, 나이가 많은 고령운전자의 사고확률이 높다는 사실을 확인할 수 있었다.



다음으로는 자동차보험의 RISK_CODE, 즉 보장내용별 변수의 영향력을 파악하고자 하였다. 이때 사용된 변수는 변수 중요도의 높은 영향을 보여주었던 하단의 2가지 변수를 사용했다.

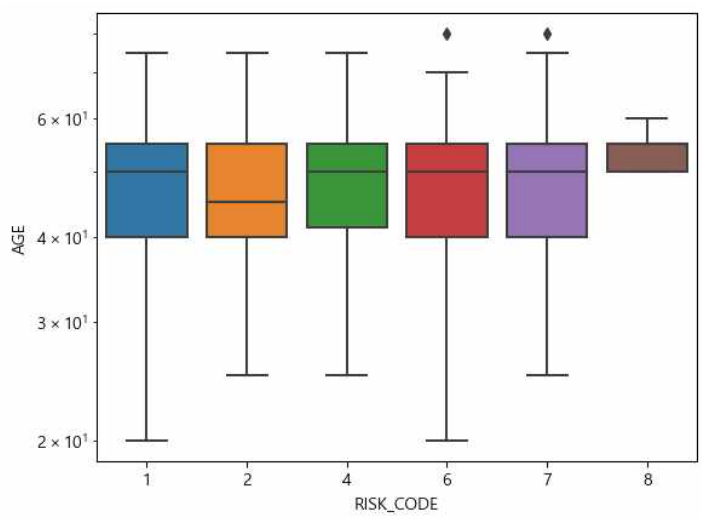
1. EARN_CNTR_CUNT 2. AGE

[EARN_CNTR_CUNT]



계약건수의 경우 무보험을 이외의 나머지 보장보험의 값은 비슷했다. 즉, 해당 보험의 경우 사고율에 영향을 미치는 계약건수 요인의 중요도를 무보험을 제외하고는 비슷하다고 유추할 수 있다. 특히 그 분포를 살펴볼 때, 대인1, 대인2, 대물, 자차에서 계약건수에 대한 분포가 넓음을 확인할 수 있다.

[AGE]



대인1과 대물 보험의 경우 나이대가 가장 다양하다. 6가지의 보험 종류 모두 나이의 중위값이 40~50대 사이임을 알 수 있다. 또한, 무보험에서 가장 나이대가 높고, 대인1, 자손, 자차의 경우 두 번째로 높음을 알 수 있다. 대인2 보험 2 보험의 경우 나이대가 가장 낮다.

이제까지의 과정을 통해 자동차보험에 대하여 사고율에 영향을 미치는 요인들에 대한 분석인 가설2를 검증할 수 있었다.

2. 기대효과 및 제언

지금껏 3가지 보험별로 사고율을 예측하고, 예측에 유의미한 영향을 주었던 요인들을 분석하였다. 또한, 예상 사고율과 실제 사고율을 정의하여 군집화 알고리즘을 통해 고객을 4가지 유형으로 분류하는 분석까지 수행하였다.

이러한 분석을 활용할 수 있는 방안은 다음과 같다.

해당 보험별 사고율에 영향을 줄 수 있는 잠재 요인을 파악하여, 초기 고객의 보험료 산정 과정에서 신상정보를 조사할 때, 집중적으로 확인해야하는 요인을 확인할 수 있다.

생명보험의 경우 중요변수였던 **미혼여부와 기간, 거주지와 직장지의 지역 일치여부, 인테리어 스타일링 관심성향, 보험계약건수**을 구체적으로 조사하며 적합한 보험료의 산출을 유도할 수 있다. 또한, 중요변수 중 하나였던 카드데이터의 총, 오후 매출건당금액을 마이데이터에서 추출하여 사고율 모델링 과정에서 효율적으로 사용할 수 있다.

또한, 생명보험의 2가지 보장 내용 중 일반사망에 대한 생명보험의 경우 미혼스코어, 오후 매출건당금액, 총 매출건당금액이 높을수록 사고를 유발하는데 더 큰 영향을 가지므로, 해당 보험에 가입하려는 고객을 대상으로 이러한 요인을 집중적으로 파악할 필요가 있을 것이다.

보험료 산출 과정은 보험의 종류마다 다르겠지만, 초기에 보험에 가입할 때뿐만 아니라, 고객이 보험에 가입해있는 기간동안 축적되는 데이터를 활용해 주기적으로 새롭게 Update 되는 상품도 많다. 따라서 생명보험에 6개의 중요변수를 파악할 수 있는 것과 같이, 해당 사고율 모델링 시 어떤 데이터에서 어떤 변수를 추출해서 가져올지 빠르게 선택할 수 있을 것이다. 또한 궁극적으로 이러한 정확한 사고율의 예측은 보험의 기본 원칙은 수지상등의 원칙에도 일맥상통하다고 말할 수 있다.

또한 이뿐만 아니라, 기존 고객의 관리 방안 역시 파악할 수 있다. 예를 들어 고객의 특성별로 프로모션이나 집중관리 등 보험사의 손실을 최대한 줄이기 위해 각기 다른 전략을 짜야될 것이다. 하지만 고객이 가지고 있는 특성은 너무 다양하다. 그렇기에 보험별 중요변수를 우선적으로 고려할 수 있다면, 고객의 지속적인 관리 전략 역시 쉽게 구축할 수 있다.

이렇듯 각 보험에 대한 '중요변수'의 핵심은 누군가에게 그 과정을 설명할 수 있다는 것이다. 본인의 보험료가 어떤 과정으로 산출되었는지, 왜 이러한 관리를 받고 있는지 쉽게 이해시킬 수 있다는 장점이 있다.

이러한 대표적인 기대효과를 바탕으로, 생명보험뿐만 아니라, 장기손해보험과 자동차보험의 분석 결과에 대해서도 생각해볼 필요가 있다.

장기손해보험의 경우 앞선 검증을 통해 보험데이터뿐만 아니라 마이데이터가 추가되었을 때 더 정확한 사고율을 예측할 수 있다는 결론을 내릴 수 있었다. 따라서 사고율 산정 시, 특정 고객에 대한 카드소비데이터를 기존 신상 데이터와 함께 이용하는 것이 보험사의 손실을 막기 위해 유리하다.

해당 보험의 경우 **보장내용, 보험계약건수, 오전 온라인쇼핑 총결제금액**의 총 3가지 변수가 사고율 예측의 중요변수라고 할 수 있다. 따라서 사전에 장기손해보험을 가입하려는 고객에 대한 정보를 파악할 때, 보험은 몇 건이나 계약이 되어 있는지, 마이데이터를 확인하여 오전 시간의 온라인쇼핑의 총결제금액은 얼마인지 파악하는 것이 사고율 예측에 큰 도움을 줄 것이다. 앞선 결과에서도 볼 수 있듯이 총금액이 많아질수록 사고확률이 미비하게 낮아지고 있음을 확인할 수 있는 점이 굉장히 특이하다. 보장내용의 경우 질병수술(입원), 상해수술(입원)의 계약이 사고율이 높다는 것을 알 수 있기에, 유관 상품을 계약할 시, 고객의 중요변수에 대한 면밀한 조사가 필요하다.

6가지의 보장내용별 요인의 영향을 분석하자면,

상해입원의 경우 계약건수,

상해수술의 경우 계약건수, 총결제금액(오전_온라인쇼핑)

질병입원의 경우 계약건수,

질병수술의 경우 계약건수,총결제금액(오전_온라인쇼핑)
급성심근경색진단의 경우 총결제금액(오전_온라인쇼핑)
를 중요하게 살펴봐야한다. 이때 명시된 변수가 높을수록 사고확률이 낮아진다.
하지만 뇌졸중진단의 경우 계약건수, 총결제금액(오전_온라인쇼핑)이 낮을수록 사고확률이 높아지기 때문에, 보험료 산출에 있어 유의할 필요가 있다.

자동차 보험의 경우
카드 데이터의 추가로 인해 오히려 사고율이 부정확하게 예측되는 결과를 유발했다. 즉 자동차보험의 사고율과 특정인의 카드데이터 사이에는 큰 연관성이 없음을 추론할 수 있다. 앞으로, 자동차 보험에 대한 사고율을 모델링할 때, 기존 보험개발원 데이터만으로도 충분히 좋은 예측을 할 수 있다는 인사이트를 준다.
해당 사고율의 경우 보장내용, 계약건수, 나이가 사고율에 영향을 끼치는 '주요변수'가 된다. 이때, 대물,자차,대인1의 경우 다른 보장 내용에 비해 상대적으로 사고율이 높기 때문에, 보험료 산정 시 주요변수에 주의하여 산정할 필요가 있다. 또한, 보험 계약건수를 확인할 때, 만약 기존 계약건수가 0.7 이상이라면 사고가 날 확률이 통계적으로 높다는 결론이 있기에, 해당 고객을 주의깊게 관리해야 한다. 나이의 경우, 일반적인 통념과 비슷한 결과가 나왔다. 나이가 어린 초보운전자와 나이가 많은 고령운전자의 사고확률이 높다는 결론을 얻었다. 따라서 현재 자동차 보험에서 실제로 시행하고 있듯이, 보험료 산출 과정에서 나이에 따른 가중치를 부과하여 적합한 보험료를 산출해야한다.
다음으로는 보장내용별 요인의 영향을 고민해보았다. 이때 계약건수 요인의 경우 무보험을 제외하고 나머지 보장보험의 사고율에 대한 영향이 비슷했다. 따라서 무보험을 제외하고 나머지 자동차 보험의 사고율을 산출 시, 공통적으로 중요하게 고려해야할 요소임을 알 수 있다. 또한, 나이 요인의 영향을 파악해보면, 대인2 보험의 경우 나이대가 낮은 고객의 보험료 산정 시, 더 많은 사고를 낼 확률이 높기 때문에 집중적으로 관리해야한다. 6가지의 보장보험의 종류 모두 대부분이 40~50대 고객임을 참고했을 때, 보험사의 홍보나, 고객의 프로모션 전략을 짤 때, 해당 나이대에 적합한 방식을 적용하는 게 중요하다. (EX. 트로트 가수 관람권 증정)

또한, 장기손해보험의 경우 카드 데이터를 활용하여 사고율을 모델링 하는 것이 더 효율적이라 결론이 났으므로 산출된 예상/실제 보험료를 바탕으로 군집화를 시도할 수 있다. 다음과 같이 고객의 유형이 분류된다.

[1] 예상 사고율이 낮고 실제 사고율도 낮은 고객

평균나이대 40~50(나이대 분포가 다양)

총 매출건당금액 (1위, 이상치가 많음)

뇌졸중진단, 급성심근경색진단

연소득 3위

[2] 예상 사고율이 높고 실제 사고율도 높은 고객

평균나이대 40~50

총 매출건당금액 (2위, 이상치가 많음)

질병수술, 질병입원 보험 계약

대부분 연소득 (6천만원 이하, 1위)

[3] 예상 사고율이 낮고 실제 사고율이 높은 고객

평균나이대 40~50

총 매출건당금액 4위

대부분 (연소득 6천만원 이하, 2위)

[4] 예상 사고율이 높고 실제 사고율이 낮은 고객

평균나이대 40~50(그 중 최고령, 나이대 분포가 다양)

총 매출건당금액 3위

연소득 4위

이때, [1] 예상사고율과 실제 사고율 모두 낮은 고객은 '우량고객'이다. 보험사의 입장에서는 가장 이상적인 유형의 고객이라고 할 수 있다. 모델이 예측을 잘 수행하고 있다는 뜻이고, 해당 고객들은 사고율이 낮아 궁극적으로 보험사가 보험금을 적게 지급해도 된다. 이러한 우량 고객들이 지속적으로 해당 보험사의 서비스를 이용할 수 있도록 "고객 유지"의 관점에서 나이대, 관심 변수, 소득 수준을 고려하여 콘서트 관람권, 가전제품 증정 등의 프로모션, 인센티브 전략을 취해야 한다.

[2] 예상사고율과 실제 사고율 모두 높은 고객은 "주의 고객"이다. 예측이 잘 이루어지고 있지만, 사고 발생 확률이 높은 고객들이기 때문에 보험사의 입장에서는 보험금을 많이 지불하게 될 수도 있다. 이러한 위험을 관리하기 위해서는 궁긍적으로 고객들의 사고율을 낮추는 방향으로 유도하는 전략이 필수적이다. 총 매출건당금액이 사고율에 많은 영향을 끼치므로 소비를 줄이는 방식을 간접적으로 유도하는 방식 또한 하나의 전략이 될 수 있다. 또한 앞서 산출한 중요변수를 사용하여 고객의 실제 사고율 산정에 가중치를 두고 모델의 재

학습을 통해 더 현실적인 보험료 산정을 가능하게 할 수 있다.

[3] 예상 사고율이 낮고 실제 사고율이 높은 고객의 경우 '이상고객'으로 집중관리의 대상이다. 보험사의 입장에서 가장 경계를 해야되는 유형이다. 보험사의 수익 구조에 손실을 가져다주는 위험 고객이기 때문이다. 따라서 해당 나이대와 연소득, 매출건당금액을 고려하여 보험료 할증 적용이나 특정 고위험 활동에 대한 보험 보장의 제한을 두는 방식을 통해 사고율 낮추는 방식을 추구해야 한다.

[4] 예상 사고율이 높고 실제 사고율이 낮은 고객의 경우 '관심 고객'이 된다. 예측 사고율이 높게 측정되기 때문에 기본적으로 고객이 지불해야 하는 보험료가 높게 산정된다 이러한 문제를 방치하면 비즈니스 관점에서 자칫 보험사에 대한 고객의 신뢰도 하락의 문제가 발생할 수 있다. 이는 장기적인 관점에서 타 보험사로의 서비스 이탈 및 중도 해지 등으로 이어질 수 있다. 따라서 해당 고객들의 보험료를 기본적으로 낮추고, 해당 40~50대에 맞는 적절한 프로모션 전략을 주기적으로 사용하는 것이 바람직하다.

3. 문제상황 및 해결 아이디어

2개의 가설을 검증하는 과정에 있어 논리적, 효과적인 분석을 위해 여러 측면에서 다양한 고민을 시도하였다. 다음은 직면한 문제상황과 그 해결 과정을 서술하였다.

(1) 사고율에 대한 정의

우선, 대회에서 주어진 사고율 공식을 사용했을 때, 여러 문제가 발생하였다. 생명보험의 경우 총 16377개의 데이터로 사고율 공식을 이루는 'ACCD_CUNT' / 'EARN_CNTR_CUNT'를 대입했을 때, 오직 7개만이 0이 아닌 값이고 나머지 16360개의 데이터는 0이있고, 나머지 10개는 결측값이었다.(분모가 0인 경우에도 결측값으로 판단) 이때, 사고율에 대한 공식이 존재하고, 결측값도 10개밖에 없는 상황에서 새로운 사고율에 대한 공식을 생성하는 것은 보험 설계 시 사고율 측정의 근본적인 오류가 될 수 있다고 판단했다. 또한, 생명보험의 PREM과 LOSS 값은 존재하였지만, 보험료와 보험금 간의 관계를 유추해서 생명보험에 대한 사고율을 구하는 것 역시 근본적인 모순이라 생각하였다. 따라서 'ACCD_CUNT' 변수에 주목하였다. 사고건수를 나타내는 해당 변수의 경우 1번의 결과론적인 사고율 공식 'ACCD_CUNT' / 'EARN_CNTR_CUNT'(초기 보험료 산정 시, 개인의 신상정보를 바탕으로 계산하는 사고율이 아닌, 실제 사고가 일어난 현황을 바탕으로 계산하는 사고율)과 가장 연관이 높다고 판단하였다. 따라서 사고율을 'ACCD_CUNT' 변수로 대체하는 방식을 사용하였다.

장기손해보험과 자동차보험의 경우 각각 14414개와 54210개의 데이터로 이루어져 있다. 이때 사고율에 대한 공식인, LOSS/PREM은 장기손해보험의 경우 18개의 결측치, 약 97%(14031/14414)가 0값을 가진다. 자동차 보험의 경우 사고율의 100% 모두 결측치이다. 두 경우 역시, 대부분의 사고율 값이 0이나 결측치를 가지는 상황에서 새로운 사고율을 대체할 수 있는 방식을 찾아야했다. 이 역시, LOSS/PREM이라는 결과론적인 사고율(보험금을 나타내는 LOSS 변수는 사고가 일어난 뒤의 사건)을 고려했을 때, 'ACCD_CUNT' 변수를 새로운 사고율로 대체하기 적합하다고 판단했다. 하지만 장기손해보험과 자동차 보험의 경우, 데이터 수가 부족한(대부분 0) 생명보험과 다르게 데이터 수가 충분하다고 판단되어 사고율으ᄅᆞᆯ 예측하는 모델링을 시도하였다는 차이가 있다. 이때, 'ACCD_CUNT' 변수를 예측하는 모델을 구성했지만, 모델링 후 고객을 군집화할 때에는 0 또는 1 범주를 예측할 때 출력된 확률값을 새로운 사고율로 대체하였다. 간소화된 모델이라 할 수 있는 로지스틱 회귀분석 모형을 통해 Classifcation 모델링 과정을 떠올려보면 시그모이드 함수에 삽입하는 x의 값이 새로운 사고율이 되는 셈이다. 일반적인 분류 모델은 CUT-OFF 값을 0.5로 설정하고 예측확률값(X)이 0.5 이상이면 1, 0.5 미만이면 0으로 분류하는데, 해당 원리에 착안한 것이다.

(2) 분석방식에 대한 정의

[1] 생명보험과 (장기손행보험, 자동차보험)의 분석 방식이 다른 이유

생명보험, 장기손해보험, 자동차보험 모두 사고유무를 바탕으로 Classification 모델링을 시도한다는 공통점이 있다. 하지만 생명보험 모형의 경우, 로지스틱 회귀 모형을 통해 변수에

대한 해석을 하는 방식이지만, 장기손해보험과 자동차보험의 경우 Tree 계열의 모형인 Xgboost를 사용해 변수를 해석하는 것뿐만이 아니다. 이에 더해 사고유무를 예측하는 확률값을 출력하여 이를 이용해 보험 데이터만을 사용해 예측한 사고율과 카드데이터를 합친 사고율을 비교한 후 군집화하여 고객을 분류한느 분석방식을 수행한다. 분석의 차이에는 생명보험의 적은 데이터 수를 원인으로 꼽을 수 있다. 생명보험에 대한 사고유무 변수를 확인했을 때, 1의 값(사고 有)을 가지는 17개의 데이터를 제외하고는 모두 0이었다. 이 경우 예측모형을 구축하기에는 턱없이 부족한 데이터의 수이며, 모델링 과정에서 중요한 '일반화'를 수행할 수 없다고 판단했다. Classification의 가장 간소화된 모델인 로지스틱 회귀분석을 사용하여 모델의 성능을 어느정도 보장하며 사고율에 영향을 끼치는 변수를 해석하는데 집중하고자 했다. 장기손해보험과 자동차봏머의 경우 1의 값을 가지는 데이터는 각각 379, 1240개로 여전히 종속변수인 Target의 범주가 Unbalanced(사고 有 대비 사고 無)라 할 수 있지만, Undersampling하여 각각의 데이터의 개수를 2배로 증가시킨다면 모델링을 수행하기 충분하다고 판단하였다. 따라서, 모델의 성능뿐만 아니라 변수의 영향을 해석하기 위한 설명력 역시 높은 Tree 기반 모델을 사용하였고, 해당 데이터의 수라면 사고율의 오차를 통해 고객을 분류할 수 있다고 판단했다.

[3] 모델을 3가지 종류로 나눈 이유

모델링 개수에 대한 고민 역시 존재하였다. 같은 생명보험이라 하더라도 일반사망, 재해사망 등 세부보장 내용에 ᄄᆞ라 사고율이 달라질 수 있고, 사고율에 영향을 주는 요인 역시 달라질 가능성이 농후하다. 3가지 보험에 대해서 'RISK_CODE'로 분류했을 때, 총 15가지의 모델을 만들 수 있다. 하지만 15개의 모델을 구축하기에는 각 모델에 대한 데이터가 부족했다. 2000~15000개 정도의 데이터로 일반적인 모델링을 수행하기에 어려움이 없으리라 판단되지만, Classification 모델링의 특성상 사고유무가 1인 데이터는 현저히 적기에 모델의 성능을 높이기 어려웠고, 낮은 성능의 모델에 대한 해석을 시도하는 과정 역시, 근본적인 모순이 있다고 판단했다. 따라서 데이터의 수를 확보하기 위해 모델을 3가지로 나눈 후, 각 모델별 'RISK_CODE'의 해석을 통해 각 보험의 세부보장별 요인의 영향을 파악하려고 시도했다.

(3) 변수 해석에 대한 고민

모델링 후 변수를 해석하는 과정에서 해석하려는 변수의 개수에 대한 고민이 들었다. 모델링에 사용된 모든 변수를 해석하는 것이 이상적이지만, 이를 종합하는 과정에서 논리적인 비약이 생길 수 밖에 없다. 예를 들어 사고율에 영향을 끼치는 100개의 모든 변수에 대해 일일이 얼마만큼의 영향을 주는지 파악하려 한다고 가정해보자. 특정 고객에 대한 보험료를 산정하고 그 과정을 고객에게 설득하는 과정에서 "변수 1의 경우는 ~ ', " 변수 55의 경우는 ~ " 등 모든 변수를 설명하려고 한다면 보험에 대한 사전지식이 없는 고객의 입장에서 전혀 이해할 수 없는 설명이 도리 것이다. 따라서 사고율에 영향을 끼치는 변수에 대해서도 '선택과 집중'은 필수적이다. 중요하게 영향을 끼치는 몇몇 변수를 파악하여 그 영향을 중심으로 고객에게 설명하고, 고객의 유형에 따른 전략을 설계했을 때, 핵심적인 내용을 쉽게 받아들일 수 있다는 장점이 있다. 따라서 본 분석의 경우, ICE와 PDP기법을 통해 변수를

해석하는 과정에 앞서, Tree 기반 모델에서 변수의 중요도를 출력하여 집중적으로 해석해야 될 변수를 정한 후 해석을 시작하는 방식을 적용했다.

[4] 모델에 대한 고민

보험산업과 같이 규제가 엄격한 산업에서 모델을 개발할 때는 일반화가 잘되는 정확한 모델을 구축하는 것이 중요하다. 하지만 이러한 모델은 모델의 결과를 설명할 수 있는 '설명력'이 낮아질수 있다는 제약이 존재한다. 따라서 이러한 설명력(Explainability)과 성능(Accuracy)의 trade-off 관계에서 두 기능을 적절하게 충족할 수 있는 모델을 선택하는 것이 중요하다. 일례로 미국의 보험위원협회(National Association of Insurance Commissioners)는 머신러닝 기반의 요율 산출모델의 평가에 사용될 3가지 기준을 제시했다. 이는 (1) 새로운 모델이 기존 보험료에 어떤 영향을 주고, 요구를 받을 시 보험사가 어떻게 설명할 것인가? (2) 모델이 제시하는, 손해율이나 비용 측정에 기여하는 리스크가 계리의 관점에서 직관적이거나 입증가능한 연관성이 있는가? (3) 모델로 인해 개인이 차별을 받지 않는가? 이다. 즉 이에 대입해본다면, 사고율 예측 모형을 구축하는 과정에서 설명가능성이 얼마나 중요한지를 알 수 있는 내용이다. 이를 기반으로 생명보험의 모델링 과정에서는 로지스틱 회귀모형, 장기손해보험과 자동차보험의 모델링 과정에서는 Tree 기반의 Xgboost 모델을 사용하였다. 로지스틱 회귀모형의 경우 Classification의 가장 단순한 선형모형이기에 총 데이터의 개수가 34개인 상황에 가장 알맞다고 판단하였다. 또한, 해당 모형의 경우 각 계수의 값과 그 통계적 유의성이 함께 출력되어, 통계적인 해석이 가능하다는 장점이 있기에 선정하였다. Xgboost 모델의 경우 Tree 기반의 앙상블 모델로 높은 성능뿐만 아니라, 작동과정에서 Impurity(불순도)의 개념을 활용하여 종속변수의 영향을 미치는 변수의 영향력을 판단할 수 있다는 장점이 있다. 따라서 여러 Tree 기반 모델 중 가장 성능이 좋았고, 설명가능성을 두루 갖춘 해당 모델을 채택하였다.

[5] 군집화 시 유형별 고객 관리 전략의 고민

앞서 보험개발원 데이털르 활용하여 산출한 사고율(A)과 카드 데이터를 함께 활용하여 산출한 사고율 간의 비교(B)를 통해 군집화를 수행했다. 이러한 과정을 통해 고객의 분류를 총 4가지로 나눌 수 있었다. 모델링 과정에서 사용된 데이터를 바탕으로 A를 통해 산출된 사고율의 경우 보험 데이터만을 사용한 **예상 사고율**이라 한다면, B를 통해 산출된 사고율은 카드 데이터가 추가됨으로써 더 정확한 사고율을 예측할 수 있는 **실제 사고율**이라고 판단할 수 있다(이때, 앞선 모델의 성능 비교에서 B모델의 예측 성능이 더 높다면 이를 검증)

따라서 예상 사고율과 실제 사고율을 비교하면 다음과 같이 4가지 유형으로 분류할 수 있다. (1) 예상 사고율과 실제 사고율이 모두 낮은 고객 (3) 예상 사고율이 높고, 실제 사고율이 낮은 고객 (4) 예상 사고율이 낮고, 실제 사고율이 높은 고객. 이러한 4가지 고객의 유형을 바탕으로 보험사의 손실을 줄이기 위한 관리 전략을 모색하는 것이 필수적이라 생각했다. 따라서 앞서 '기대효과 및 제언' 목차에서 설명한 고객 관리 전략에서 볼 수 있듯이, 다양한 관점에서 고객을 어떻게 관리해야하는지 고민할 수 있었다. 보험사의 입장에서 단순히 가장 큰 손실을 안겨줄 수 있는, 예상 사고율은 낮으나 실제 사고율이 높은 고객에 대해서만 관리 전략을 구축하는 것이 아니라 다른 유형의 고객에 대해서도 지속적이고 효과적인

관리가 필요하다는 것을 깨달았던 부분이다.

4. 느낀점

(1) 해당 산업의 실제 데이터를 이용해 분석을 수행하면서 데이터의 처리방식과 분석 방식에 대해 다양하게 고민할 수 있었다. 대학생의 입장으로서 평소 데이터 분석에 대한 프로젝트나 공부를 진행할 때, 대부분의 데이터는 결측치가 많이 없고, 비교적 많은 내용이 담겨있는 형태였다. 따라서 별다른 처리 없이 간단한 전처리를 이용해 EDA를 하거나, 모델에 적용하는 방식으로 데이터 분석을 진행하였다. 하지만 해당 대회에서 제공받은 현업 데이터의 경우, 자동차보험의 사고율 공식에 적용하여 사고율을 구해보면 모든 데이터의 값이 결측치일 정도로 결측치가 굉장히 많았다. 따라서 실제로 예선계획서에 작성했던 방식을 비교적 동일하게 적용하기 위한 전략에 대해 여러번 고민하였다. 예를 들어, 사고율을 대체할 수 있는 분석방법에 대해서도 다음과 같이 여러 대안을 생각해보았다.

[1] PREM 값에 의존하지 않는 사고율을 새로운 공식으로 정의하기
[2) 사고율을 예측하는 모델링이 아닌 결측치가 없는 변수를 사용해 비지도학습의 일종인 군집화를 수행하며 고객의 특성을 분석
[3] 기존 사고율 공식을 적용하고 모델링을 진행하되, 공개된 실제 사고율 통계와 비교

등 기존에 설정한 분석 목적을 달성하기 위한 대체방식을 생각하고 논리적으로 오류가 없는지 계속해서 확인하는 과정을 반복하였다. 그 과정에 있어 어떤 방법을 채택할 때마다 '왜'라고 질문하며 그 근거를 충실히 세우고자 노력하였다. 그 결과, 처음에는 대회의 주제가 사고율의 분석임에도 사고율을 구할 수 없는 상황에 막막했으나, 이를 대체할 수 있는 여러 대안을 생각하고 각각 검증하는 과정을 통해 본 분석의 과정과 결과에 대한 자신감이 생기게 되었다. 후에 데이터분석가를 꿈꾸는 학생으로서, 실제 현업에서 데이터분석을 수행할 때는 현재보다 더욱 처리하기 어려운 데이터를 다룰 확률이 매우 높다. 이 경우에도, 이번 대회의 경험을 살려 여러 대안을 생각해내며 분석의 목적에 맞는 성공적인 결과를 도출할 수 있을 것 같다.

(2) 보험과 카드에 대한 데이터를 다뤄보면서 금융 산업, 특히 보험사의 상황에 대한 여러 지식을 습득할 수 있었다. 기존에는 보험사의 경우, 보험료와 보험금의 차액으로 수익을 창출하는 비즈니스 모델을 가지고 있음을 추측하였다. 하지만 다양한 자료를 찾아보면서 그 차액으로 수익을 얻을 수는 있겠지만, 수지상등의 원칙에 부합하도록 정확한 사고율에 기반하여 보험료와 보험금의 총액이 같도록 설계하는 것이 일반적인 보험료 산출 과정임을 파악할 수 있었다. 이러한 배경지식을 습득하고 난 뒤, 특정 고객에 대해 더 정확한 사고율을 측정하는 것이 보험사의 손실을 방지하는 측면에서 얼마나 중요한 것인지 이해할 수 있었던 것 같다. 또한, 모델링의 과정에서도 보험사의 입장에서 알맞은 모델을 선택해야 했다. 인공지능 경연 대회 플랫폼인 Dacon이나 Kaggle 등에서는 모델의 설명 가능성보다는 모델의 성능에 초점을 둔다. 이는 모델의 작동방식에 대한 해석보다도 그 성능이 좋다면 더 훌륭한 분석임을 보여준다. 하지만 보험사의 입장에서 보험료의 기반이 되는 사고율을 측정하고, 이를 고객에게 설득시키고, 실제로 고객을 관리할 때의 전략을 파악하는 과정 등에서 모델

을 해석하는 과정이 필수적이다. 스스로가 왜 해당 모델이 좋은 성능을 가지는지, 어떤 변수가 그 성능에 영향을 미쳤는지 파악할 수 없다면, 아무리 사고율을 정확하게 예측해도 신뢰할 수 없는 모델이 되는 것이다. 이처럼 산업의 분야에 따라 모델링을 할 때, 고려하는 요소가 분석목적에 따라 다르다는 것을 체감할 수 있었다. 설명가능성과 성능이라는 두 마리 토끼를 잡아야하는 보험사의 입장을 고려해보면서, 이를 충족시킬 수 있는 XAI에 대한 많은 공부를 할 수 있었다.

(3) 데이터 분석 결과를 제시하는 과정에서 실제 보험사의 현업자의 입장에서 생각해보며 단순히 결과를 보여주는 분석이 아니라, 그 과정과 결과에 따른 활용방안을 논리적으로 설득하는 것이 얼마나 중요한지 파악할 수 있었다. 단순히 PDP 방식을 통해 '사고율에 어떤 변수가 어느정도 영향을 끼친다' 는 해석에서 나아가 초기 보험료 산출 시 개인 신상에 대한 정보를 수집할 때, 해당 변수의 내용을 더욱 집중적으로 수집하는 방식과 같은 전략을 고민하기도 하였다. 또한, 고객 데이터의 지속적인 Update를 통해 특정 기간별 사고율이 변경될 것이고, 이를 통해 고객에 대한 CRM 전략을 새롭게 짤 수 있을 것이다. 즉, 사후분석의 영역 역시 중요함을 깨달았다.

또한, 데이터 분석의 과정에 있어 코딩뿐만 아니라 분석 아이디어를 수립하는 과정 역시 중요하다는 것을 느꼈다. 분석목적을 설정하고, 목적을 달성하기 위한 여러 분석 방법을 도출하면서 동일한 목적에 대해서도 이를 논리적으로 표현할 수 있는 여러 참신한 아이디어의 중요성에 대해 깨달았던 것 같다. 분석 목적을 설정하고, 여러 분석방법론에 대한 아이디어를 고민하고, 실제 코드를 통해 이를 구현하며 효과성을 분석,해석하는 과정을 수행한 데이터 분석가가 되기 위해서 얼마나 많은 공부와 경험을 해야하는지 스스로 되돌아볼 수 있었던 시간이기도 하였다. 추가적으로, 분석의 결과를 다른 사람에게 제시하는 효과적인 방식을 고민해야 한다고 느꼈다. 보고서나 소스코드를 다른 사람에게 보여줄 때, 시각화를 통해 결과를 효과적으로 보여주며, 코드의 경우에도 코드를 간결하게 제시하는 것뿐만 아니라 목차나 주석을 제시하는 방식을 통해, 그 내용과 더불어 상대방이 쉽게 이해할 수 있게끔 작성하는 것 역시 분석가의 핵심적인 역량임을 느낄 수 있었다.

5, 참고문헌 및 분석환경

(1) 참고문헌

[1] 허경옥, 박상미, 박귀영.(2011). 가계의 보험보유량 및 보험료지출의 영향요인 및 보험료지출의 적정도 분석.한국FP학회지,4(1),159-182.

[2] 김경자, 홍정하. (2002). 우리나라 가계의 생명보험료 제출실태와 그 영향요인.대한가정학회 학술대회,(),91-91.

[3] 정중영, 강중철.(2006). 자동차보험 손해율에 관한 연구. Journal of The Korean Data Analysis Society,8(6),2445-2456.

[4] 임선호, 박은미, 장현봉.(2009). 교통사고율에 영향을 미치는 요인 분석.대한교통학회지,27(4),41-53.

[5] 오창수, 문성철. (2012). 장기손해보험의 보험리스크 산출에 관한 연구. 계리학 연구, 4(2), 17-46.

(2) 분석환경

pandas	1.4.4
seaborn	0.11.2
numpy	1.21.5
matplotlib	3.5.2
xgboost	1.7.2
sklearn	1.0.2
tqdm	4.64.1