별첨2 **데이터 분석·모델링 개발 계획서(안)**

□ 개요

1. 요약

기존에는 보험료를 산출하는 과정에 있어 보험사고의 발생률, 즉 사고율 산출 시 한정된 개인의 인적정보만으로 판단할 수밖에 없었다. 하지만 각 개인의 소비패턴을 보여주는 카드 명세서 등의 마이데이터를 사고율의 산출과정에서 고려할 수 있다면. 개인의 특성을 잘 반영할 수 있는 합리적, 객관적인 보험료를 산출할 수 있다. 본 분석에서는 고객의 카드 명세서를 활용하여 보험료 산출의 기반이 될 사고율을 새롭게 정의(예측)하고, 해당 사고율에 영향을 주는 변수에 대해 분석하고자 한다. 또한, (1)카드 명세서 데이터를 고려하지 않고 기존 보험명세서 데이터만을 고려하여 사고율을 계산한 경우, (2)카드 명세서 데이터와 기존 보험명세서 데이터를 모두 고려하여 사고율을 계산한 경우, 이 2가지 경우에 대해 각각 군집화 알고리즘을 사용하여 고객의 특성을 4가지로 분류할 것이다. 그 후, 각각의 군집에 대한 특성을 알아본 후, 이상 고객(보험사가 예측한 사고율과 실제 사고율의 괴리가 큰 고객)이 포함된 군집을 판별하여 본 분석의 핵심이라 할 수 있는 (2)번 경우의 효과성을 알아볼 것이다. 그 효과가 입증된다면, 이상 고객의 패턴을 분석하며 보험사가 고려하지 못한 요인, 즉 합리적인 보험료 산출을 위해 보험사가 집중적으로 고려해야 할 요인을 분석하고자 한다.

2. 주요 방법론

카드 명세서를 기반으로 새로운 사고율을 산출하는 과정에 사용되는 예측 모델링, 최적화된 예측 모델을 바탕으로 종속변수인 사고율에 영향을 끼치는 변수에 대한 해석, 앞서 언급한 (1)번과 (2)번 경우에 대해 예측한 사고율과 실제 사고율을 기반으로 고객을 4가지 유형으로 분류하는 군집화 알고리즘, 각 군집에 대한 EDA를 바탕으로 이상 고객을 판단한 후 (2)번 방식에 대한 효과 검증, 이상 고객에 대한 패턴을 분석하는 요인분석을 주요 방법론이라 할 수 있다.

예측 모델링의 경우, 3가지의 보험 종류별로 데이터를 나누어 진행할 것이다. 이때, 데이터의 크기가 모델을 검증할 수 있을만큼 충분히 크다면, 15가지의 보험 세부내역을 기준으로 모델링을 진행할 것이다. 결측치, 이상치, 스케일링(Scaling)과 같은 기본적인 전처리를 진행한 후, Factor analysis, PCA 등의 차원축소 기법, Tree 기반 모델의 변수 간 상호작용 파악한다. 이를 바탕으로 파생변수 생성 및 불필요한 변수 제거를 통해 모델링에 사용되는 변수의 개수를 줄일 것이다. 또한, 데이터 불균형(사고율이 0인 데이터 多)을 상쇄하기 위해 소수나 다수의 데이터를 바탕으로 데이터를 증가, 감소시키는 Oversampling 및 Undersampling 기법을 적용하려 한다. 그 후 교차검증 기법을 사용하여 모델을 최적화할 것이다. 이때 예측 모델에는 Xgboost, Lightgbm, Catboost, Randomforest와 같은 Tree 기반 앙상블 모델을 사용한다.

변수 해석에 있어서는 Tree 기반 모델을 바탕으로 한 변수중요도를 사용하고, Partial

Dependence Plot (PDP) 그래프를 통해 Marginal contribution을 파악할 것이다. 또한, 각 변수에 대한 shapley value 기반의 영향 분석 역시 실시하려 한다.

군집화 알고리즘의 경우, 데이터의 분포가 고르지 않다는 점을 예상하여 밀도를 기반으로 군집화를 수행하는 (H)DBSCAN을 사용할 것이다. 이때, 군집화의 결과에 따라 데이터 스케일링(Scaling) 같은 추가적인 처리를 고려할 수 있다.

이상 고객에 대한 효과 검증의 경우, 각 군집에 대한 EDA를 통해 군집의 특성에 대해 파악한 후, 이를 기반으로 이상 고객이 속한 집단을 선정한다. 그 후, 앞선 (1)번과 (2)번 경우에 대한 이상 고객의 절대적 수치를 비교하여 그 효과를 검증할 것이다.

이상 고객에 대한 패턴을 분석하는 요인분석의 경우, 앞선 변수 해석에서 사용한 Partial Dependence Plot (PDP) 그래프 및 shapley value 기반의 영향 분석을 통해 이상 고객에 대한 주요 변수의 특징을 파악하려고 한다.

3. 분석・모델링 기법 선택 배경

분석 기법의 경우, 핵심은 카드 명세서라는 개별 고객에 대한 마이데이터 정보를 추가해 사고율을 새롭게 예측하고, 그 전후 효과의 개선을 입증하는 것으로 판단했다. 이에 더해, 사고율에 영향을 주는 변수의 구체적인 효과를 파악하여 고객의 보험료 산정과정에서 사용될 새로운 전략을 모색하는 것이 필수적일 것이다.

따라서 우선, 카드 명세서를 바탕으로 사고율에 영향을 줄 후보 변수들을 생성하여 더 객관적으로 사고율을 예측하는 모델을 만들고, 영향력 있는 변수를 사람이 주관적으로 판단하는 것이 아니라, 수리적 알고리즘을 통해 모델이 스스로 판단하는 방식이 합리적이라 생각했다. 또한, 이러한 방식에 대한 효과성을 입증하기 위한 도구로 실제 사고율과 예측 사고율을 통해 고객의 유형을 분류(군집화)하고 보험사가 집중적으로 관리해야 할 고객의 군집(=이상 고객이 포함된 군집)을 찾고 그 군집 내 개체 수를 비교하는 방식을 사용하였다. 만약, 카드 데이터를 포함하지 않은 기존의 데이터를 통해 산출된 이상 고객의 수보다, 카드 데이터를 포함한 후 산출된 이상 고객의 수가 적다면 사고율을 정확히 예측하는 것에 효과가 있다고 판단할 수 있다. 이뿐만 아니라, 보험사의 입장에서 이상 고객의 분류에 많은 영향을 끼치는 변수(요인)을 확인할 수 있다면 보험료를 재산정시, 혹은 새로운 고객에 대한 보험료를 산정 시, 이를 고려하여 더 합리적인 보험료 계산이 가능하다고 판단했다.

모델링 기법의 경우, 만약 데이터의 크기가 충분하다면, 15가지의 보험 세부내역을 기준으로 데이터를 나누어 각각 모델링을 진행(총 15번)하는 것이 사고율에 대한 더 정확한 예측과 변수에 대한 해석을 할 수 있을 것이다.

또한, 파생변수 생성 및 불필요한 변수를 제거하는 방식은 모델링에 사용될 변수의 차원을 축소하는 과정이다. 이 경우 기존 데이터의 많은 변수로 인해 발생할 수 있는 과적합을 방지할 수 있고, 변수의 영향을 해석하는 과정에 있어서도 더욱 수월하다.

데이터 불균형의 해소 방식의 경우, 실제 데이터를 살펴보면 '사고건수' 변수가 0인, 즉 모델링의 종속변수라 할 수 있는 '사고율'이 0인 변수가 많을 것이다. 이를 처리하지 않고 모델링 시, 손실 함수를 최소화할 수 있는 최적의 모델을 찾는 것을 방해하게 된다. 따라서 Undersampling이나 Oversampling으로 데이터의 균형을 맞추어 사고율을 잘 예측할 수 있는 최적의 모델을 만들 수 있다.

구체적인 모델의 경우 Xgboost, Lightgbm 등의 Tree 기반 앙상블 모델을 사용하는데, 이는 모델의 설명가능성을 고려한 방식이다. 보험료를 산정하고 고객에게 이를 설명할 때, 사

고율은 어떻게 산정되었고, 사고율에 영향을 끼치는 변수의 효과를 정확히 설명하는 것이 중요하다. 딥러닝과 같은 타 모델의 경우 모델링 결과에 대한 정보를 바탕으로 고객을 설득하기에는 설명 가능성이 부족하다는 단점이 있기에 본 분석에서 제외하였다.

변수의 영향 분석에서는 Tree 기반 모델의 변수중요도와 Partial Dependence Plot (PDP), shapley value 기반의 영향 분석을 사용하였다. 이는 각각 사고율을 예측할 시 독립변수 중 유의미한 정보가 많았던 변수를 파악하고, 이러한 변수가 어떤 수치적인 영향을 끼쳤는지, 각각의 고객 개체에 해당하는 그 수치적 영향은 어떠한지 확인하는 과정으로 변수의 영향을 총체적으로 파악하기 위해 선정하였다.

군집 분석의 경우 (H)DBSCAN을 사용할 것이다. 이는 밀도 기반의 군집화 방식으로 데이터의 분포에 기반해 다양한 모양으로 군집을 형성한다. 사고율 데이터의 특성상 그 분포가 고르지 않다는 것을 고려하여 거리 기반으로 원 모양의 군집을 형성하는 K-Means 방식보다 더 효율적이라 판단하였다.

4. 기대효과

첫째, 각 고객의 소비패턴에 기반한 사고율을 정확하게 예측하여 객관적, 합리적인 적정 보험료를 산출할 수 있다. 개인의 일부 인적사항으로 사고율을 판단하는 기존 계산 방식의 경우, 개인의 세부 특성을 고려하지 않으므로 보험료를 객관적으로 파악하는데 확률적인 제약이 존재한다. 하지만, 마이데이터와 같은 개인의 특성이 추가된다면, 이를 수리적인 알고리즘을 통해 개인의 모든 특성을 고려한 보험료를 산출할 수 있게 되고, 곧 수지상등의 원칙에 부합한다. 적정한 보험료의 산출은 많은 보험계약자의 이익을 보호함과 동시에 보험자 간 지나친 경쟁을 억제하며 보험경영을 보호할 수 있게 도와준다.

둘째, 사고율에 영향을 주는 주요 요인을 파악하며 고객의 신뢰를 형성할 수 있다. 고객에게 설정된 보험상품, 보험료에 대한 객관적인 수치를 이해하며 그들을 설득시키는 과정을 통해 자연스럽게 해당 보험사에 대한 신뢰를 높이고, 중도 계약 해지율 역시 감소할 가능성이 높다. 정확한 데이터를 기반으로 산정된 보험료임을 보여준다면 보험 산업에 대한 소비자의 부정적인 인식 역시 변화할 것이다.

또한, 보험사의 입장에서 각 보험의 세부내역별 종류에 따라 사고율에 집중적으로 영향을 끼치는 요인을 찾게 된다면, 해당 요인에 대한 가중치를 주는 방식 등의 집중적 관리로 더 합리적으로 보험료를 산출하고 고객을 관리할 수 있다.

셋째, 실제 사고율과 예상 사고율을 바탕으로 고객의 유형을 분류하여 고객에 대한 맞춤형 관리 전략을 세울 수 있다. 각 고객의 유형의 특징을 파악하면서, 보험료 변경, 보험 관련 프로그램 홍보 등 각 유형에 적합한 고객 관리 방안을 모색할 수 있다. 또한, 이상 고객의 유형을 식별하고 그 패턴을 분석하면서 보험사가 간과했던 변수(요인)에 대해 자세히 파악할 수 있다. 이를 통해 해당 변수를 사용하여 고객의 실제 사고율 산정에 가중치를 줌으로써 새로운 보험료 산정 및 새로운 고객의 보험료 산정을 위해 활용될 수 있다.

□ 참가팀의 핵심 기술 설명

본 모델의 핵심은 마이데이터에서 제공되는 카드 거래 내역을 바탕으로 새로운 고객 데이터가 주어졌을 때 사고율을 기존보다 더 정교하게 예측하는 데 있다. 전반적인 예측력을 향상하고, 모델을 참고하여 실무진들이 비즈니스 관점에서 맥락에 맞는 의사결정을 내릴 수 있도록 다양한 전처리 및 머신러닝/데이터마이닝 분석 기법을 시도할 계획이다.

1. 데이터 전처리 과정에서 사용된 분석기법

우선 feature(변수) engineering이다. 기존 변수들의 활용만으로는 보다 더 정교한 사고율 예측을 수행하기 어렵다. 따라서 건강, 소득, 시간대 등 다양한 관점에서 비즈니스 맥락에 맞는 새로운 파생변수를 형성하여 모델이 feature를 다채롭게 학습할 수 있도록 한다. 해당 과정은 또한 추후 모델링에 의한 결과 해석 과정에서도 큰 도움이 될 수 있다. 그리고 과적합과 차원의 저주(curse of dimensionality)의 가능성을 우려하여 새로운 feature가 만들어진 후, 사용되었던 기존 변수들은 제거하는 것이 중요하다.

다음은 활용될 수 있는 신규 파생변수의 예시다:

1) 업종별 소비율
= 소득 대비 업종별 매출 금액
고객이 어느 부문에서 소비 성향이 높은지, 자산 대비 과소비를 하는 경향은 없는지 확인할 수 있게 해준다.

2) 업종별 매출 건수 비율
= 전체 매출 건수 대비 업종별 매출 건수
고객의 업종별 소비 빈도(Frequency)를 확인할 수 있게 해준다.

3) 업종별 매출 금액 비율
= 전체 매출 금액 대비 업종별 매출 금액
고객이 어느 부문에서 지출이 많은지 업종 간 상대적 비교가 가능하다.

4) 시간대별 매출 건수 비율
= 전체 매출 건수 대비 시간대별 매출 건수
어느 시간대에 고객의 카드 사용이 활발한지 빈도를 확인할 수 있다.

5) 시간대별 매출 금액 비율
= 전체 매출 금액 대비 시간대별 매출 금액
어느 시간대에 고객의 정량적인 지출이 높은지 확인할 수 있다.

이외에도 데이터셋이 주어지면 변수 간 상관성을 분석하여 상관성이 높은 변수끼리 묶는

방식으로 파생변수를 생성할 수 있게 된다. 활용할 수 있는 방법론으로는 factor analysis, PCA, 혹은 Tree 기반 모델에서 변수 간의 상호작용을 고려하여 묶는 방법 등이 있고, 다양한 방법을 시도할 예정이다.

추가적으로 feature elimination도 고려할 수도 있다. 변수 중 대분류뿐만 아니라 중분류, 소분류 지표가 세부적으로 존재할 경우, 더 구체화된 형태의 변수만을 사용하기 위해 노력하였다. 가령 소분류 "매출건수합계_의료기기용품"의 경우 [건강/건강보조/의료기기용품] 항목에 속하기 때문에 중분류 "매출건수_건강보조"와 대분류 "매출건수_건강" 변수는 사용이 불필요하다. 이는 데이터 변수 간 서로 중복되는 의미를 갖게 되는 경우, 추후 분석 과정에 있어서 다중공산성 문제를 일으킬 수 있기 때문에 조치하는 것이 필요하다.

또한, 사고율을 예측하는 과정에서 데이터 불균형 문제가 필연적으로 생긴다. 통상적으로 사고에 해당하는 데이터 수는 그렇지 않은 정상 건수(=사고율이 0인 데이터)에 비해 현저히 적다. (ex) 건강한 고객들이 그렇지 못한 고객들보다 당연히 훨씬 많을 것이고, 자동차 보험의 경우도 교통사고를 일으키는 고객들보다 사고가 나지 않는 고객들이 통상적으로 많다.) 그러나 종속변수(Y)에 해당하는 데이터의 수가 다른 유형에 비해 지나치게 적은 상황이 발생하면 추후 데이터 분석 과정 및 모델링 결과와 성능 해석에 있어 다양한 문제가 발생한다. 그렇게 되면 필연적으로 우리 모델은 불균형 학습을 수행하여 잘못된 해석 결과를 제공하게 되는 치명적인 단점이 나타난다. 따라서 정확한 데이터 분석을 위해 데이터 Resampling 과정은 필수로 수행되어야 한다. 이때 주류 데이터를 조절하여 소수 데이터 수와 동일하게 만드는 Undersampling 기법이나, 반대로 소수 데이터의 샘플을 복제하여 전체 데이터 간의 균형을 맞추는 SMOTE 등의 Oversamping 기법을 활용할 것이다.

2. 모델링 과정에서 사용된 알고리즘

모델링에 있어서 핵심적으로 고려한 사항은 Explainable AI의 관점인 "설명가능성"이다. 아무리 성능이 좋은 모델일지라도 그것이 비즈니스 관점으로 어떤 함의가 있는지, 그리고 해당 결과를 도출하기 위해 구체적으로 어떤 과정을 수행해야 하는지 모델이 말해주지 못하면 무용지물이다. 이는 금융, 특히 보험 영역에서 중대한 사안이다. 보험사의 관점에 입각해, 1) 어떤 feature가 중요한 영향력을 끼쳤는지, 2) 예측한 사고율이 실제 값과 차이가 난다면 (예측에 실패했다면) 그 이유가 어떤 요인에 의해 발생했는가를 설명 할수 있는지, 3) 모델이 보험사의 수익성에 기인하는지 (보험사에게 돈을 벌어다 줄 수 있는 비즈니스 구조이고, 손실의 위험은 없는지) 등을 중시하며 모델링을 수행할 계획이다.
"설명가능함"을 중시하는 관점에서 Tree기반의 모형들을 사용할 것이다. 의사결정나무를 뿌리로 두고 있는 XGBoost, LightGBM, RandomForest, 그리고 CatBoost 총 4가지 모델을 활용하고자 한다. 의사결정나무에 입각한 모델을 설명한 이유는, 모델링 시 연속적으로 발생하는 의사결정 문제를 시각화하여 의사결정이 언제, 어떻게 이뤄지는지 성과를 한눈에 볼 수 있어서다. 즉, 모델의 계산 결과가 직접 나타나기 때문에 "해석 가능"하다는 것이다. 가령 특정 고객에 대한 사고 확률이 기존의 주어진 방식대로 산출한 것과는 다르게 높게 나왔다고 하자. 이때 해당 고객이 정말 위험 고객인지 판별하기 위해 어떤 논리에 의하여 다음과 같은 결과가 나왔는지 이유를 설명할 수 있는 "해석력"을 우리 모델은 지니고 있을 것이다.

추가적으로 특정 단일 모델을 활용했을 때, 데이터에 과적합되어 좋은 성능이 나오지 못할 수도 있다. 이러한 과적합 문제를 해결하기 위해 여러 모델의 결과를 혼합하는 앙상블(Ensembling)도 적용할 것이다. 핵심 아이디어는 예측력이 약한 모형들을 결합하여 강한 하나의 예측 모형을 만드는 것이다, 이를 통해 여러 종류의 단일 모델들 간의 장단점을 보완해주고 성능이 안정적으로 나오는 것을 확인할 수 있을 것이다. 그리고 K-fold Cross Validation(교차 검증)을 수행하여 모델을 최적화시키기 위해 노력할 것이다.

3. 결과 해석 시 활용되는 방법론

해석은 비즈니스 인사이트 도출을 위한 가장 중요한 단계이다. 모델의 결과가 도출되면, "설명가능성"을 염두하여 구체적으로 예측값이 나오기까지 어떤 변수들이 영향을 끼쳤는지 파악을 할 필요가 있다. 활용될 방법론은 총 3가지다:

1) Feature Importance

독립변수 중에서 종속변수(사고율) 예측할 때 유의미한 정보가 가장 많이 포함되었던 변수들을 순차적으로 파악할 수 있는 시각화 기법이다. 이는 Tree 기반의 모델을 앞서 활용하였기 때문에 도출할 수 있는 결과다.

모델에 의해 계산된 특성 중요도를 나타내는 수치를 막대 그래프로 알아보기 쉽게 표시할 수 있다. 이때, 각 특성에 대한 중요도는 해당 특성이 예측 모델의 결과에 얼마나 영향을 미치는지를 나타낸다. 즉, 예측에 중요한 특성을 식별하는 데 큰 도움이 된다. 또 반대로 예측에 큰 영향을 주지 않는 불필요한 변수도 확인하여 추후 개선 작업을 수행할 수 있다.

2) Partial Dependency Plot

모델에서 특정한 feature가 최종 사고율 예측에 어떤 수치적인 영향(Marginal Contribution)을 끼쳤는지 시각적으로 나타낼 수 있다. 기존 Feature importance와의 차이점은 특정 feature 값이 변할 때 모델의 예측 결과가 어떻게 변화하는지를 보여준다는 것이다. 경우에 따라 특정 feature와 예측 결과 간의 비선형 관계를 파악할 수 있도록 도와준다.

3) Shapley Value 기반 영향 분석

Shapley value는 그룹 내 각 개체의 기여도를 공정하게 분배하는 것을 목표로 한다. 모델링을 통해 해당 개념을 활용하여 변수 간의 상호작용을 평가하고, 각 변수가 예측 결과에 기여하는 정도를 추정한다. 이때 차이점은 고객 개인(개체)별로 각각의 변수가 사고율 예측에 어떤 수치적인 영향을 끼쳤는지 파악할 수 있게 해준다.

4. 페르소나 선정을 위한 방법론

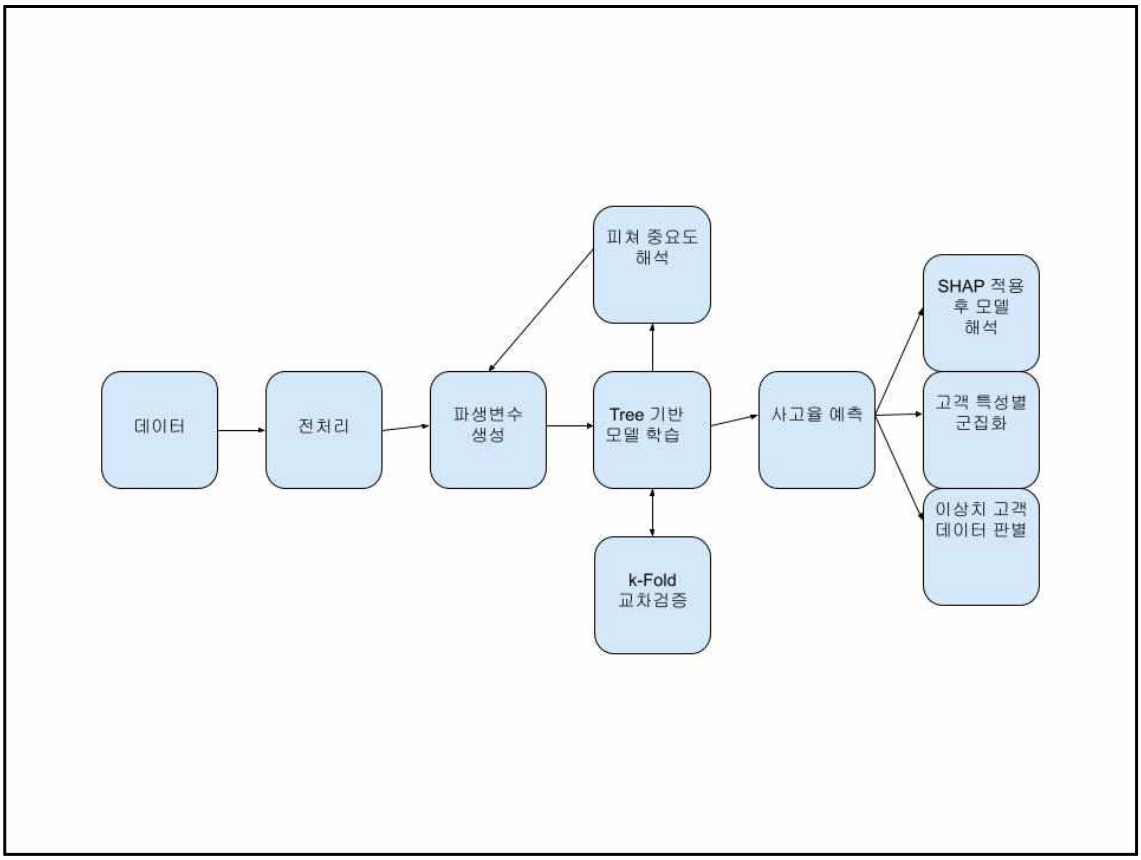
본 시스템에서는 신규 보험 가입 고객이 발생 시 이에 대한 효과적인 고객 관리를 수행하기 위해 총 4가지의 유형으로 "고객 페르소나"를 구분하고자 한다. 각 페르소나 유형에 대한 세부적인 내용은 추후 본 제안서의 "예상결과" 부분에서 더 자세히 서술하였다.

이때 페르소나에 대한 선정 기준은 기존의 실제 사고율과 해당 모델이 예측한 신규 사고율 간의 비교 분석을 통해 이루어진다. 실제 사고율 그리고 예측된 사고율의 수준이 높고 낮음에 따라 구분하여 총 4가지로 분류하고자 한다.

4가지 페르소나를 형성하기 군집화를 진행한다. 이때 알고리즘의 경우, 데이터의 분포가 고르지 않다는 점을 예상하여 거리 기반으로 원 모양의 고정된 군집을 형성하는 K-Means 방식보다 밀도를 기반으로 군집화를 수행하는 (H)DBSCAN을 사용할 것이다. 이는 밀도 기반의 군집화 방식으로 데이터의 분포에 기반해 다양한 모양으로 군집을 형성한다. 또한 형성된 군집들은 반드시 비즈니스 맥락에 맞게 의미를 부여하여 페르소나로 산정할 것이다.

□ PoC(Proof of Concept) 프로그램 설명

ㅇ 데이터 명세서와 샘플 기반으로 만든 PoC 프로그램



[그림 1] 분석 과정도

해당 과정도는 분석의 흐름을 보여준다. 구체적인 분석 방법은 하단에 서술할 것이다.

(1) 사고율 예측 모델링

분석에 앞서 과제에서 제시한 샘플 데이터를 활용하였다.

보험 종류별로 사고율 공식이 상이한 것을 고려하여, 각 종류별로 데이터를 분리하여 모델링을 적용할 것이다. 만약, 실제 데이터의 크기가 충분히 크다면, 15가지의 보험 세부내역을 기준으로 모델링을 수행할 것이다.

```
# 보험종류별 분할
df_life = df[df['보험종류'] ==1].reset_index(drop=True)
df_damage = df[df['보험종류'] ==2].reset_index(drop=True)
df_car = df[df['보험종류'] ==3].reset_index(drop=True)

# 사고율 공식
df_life['사고율'] = df_life['사고건수'] / df_life['계약건수']
df_damage['사고율'] = df_damage['지급금액'] / df_damage['보험료']
df_car['사고율'] = df_car['지급금액'] / df_car['보험료']
```

'기준년도'나 '보험종류' 등 모델이 종속변수를 예측하는데 의미있는 정보를 제공하지 않거나, 모델을 사용할 때 의미가 없는 변수를 제거하는 과정이다. 추가적으로 '계약건수', '보험료', '사고건수', '지급금액' 과 같이 예측하고자 하는 종속변수('사고율')에 대한 직접적인 정보를 노출시키는 변수 역시 제거한다. 실제 데이터에 적용 시, Factor Analysis나 PCA 등의 차원축소 기법이나 모델 기반의 상호작용 파악을 통한 파생변수를 생성하여 변수의 개수를 줄이는 방식 역시 고려할 것이다. 이러한 과정은 과적합(Overfitting)을 방지하고 모델의 정확도를 높이며, 후에 변수 해석 시 용이하다.

```
# 변수 제거
df_life =df_life.drop(['기준년도', '보험종류', '계약건수', '보험료', '사고건수', '지급금액'], axis=1)
```

모델의 학습에 앞서, 범주형 데이터의 경우 Catboost 알고리즘이 자동으로 처리할 수 있게 변수명을 지정하였다. 세부 모델은 Xgboost, Lightgbm, Random forest와 같은 종속변수의 영향을 끼치며 변수의 효과를 '설명할 수 있는' 모델을 사용할 것이다.

```
# 범주형 자료의 변수명 지정
life_cat_col = [
    '보장내용', 'JOB', 'INCOME', 'HOM_MGPO', 'HOM_SGG',
    'OFFI_MGPO', 'OFFI_SGG', 'EST_LFSTG']
```

또한, 데이터의 불균형(사고율이 0인 데이터가 많을 것으로 예상)을 상쇄하기 위해 Oversampling이나 Undersampling 기법을 적용할 것이다.

하단의 코드는 5-fold-cross validation(교차검증)을 적용한 Catboost 회귀 모델이다. 이때 교차검증이란 데이터를 여러 번 분할하여 여러 모델을 학습시키고 평균적인 성능을 계산하는 방법이다. 이는 모델의 일반화된 성능을 평가할 수 있기에 최적화된 모델을 도출할 수 있게 도와준다. 따라서 데이터를 8:2 비율로 학습 데이터(Train)와 검증 데이터(Test)로 분할 후, 학습 데이터(Train)를 다시 5분할 하여 교차검증을 진행할 것이다. 모델의 성능을 평가하는 지표의 경우 RMSE를 사용하였다. RMSE의 경우 예측값과 실제값 사이의 차이를 측정하는데 사용되는 측정 지표로, 모델이 이상치를 얼마나 잘 처리하는지 평가할 수 있으며 모델 간의 성능을 비교할 때 용이하다는 장점이 있다.

```
# 교차검증을 위한 코드
X_train, X_test, y_train, y_test = train_test_split(df_life.drop(['사고율'], axis=1), df_life['사고율'])
n_splits =5
cv = KFold(n_splits=n_splits, shuffle=True)
scores = []
models = []
for train_idx, val_idx in cv.split(X_train):
    model = catboost.CatBoostRegressor()
    model.fit(
        X_train.iloc[train_idx], y_train[train_idx],
        eval_set=[(X_train.iloc[val_idx], y_train[val_idx])]
    )
    models.append(model)
    scores.append(model.get_best_score()["validation"]["RMSE"])
```

하단은 교차검증을 통해 설정된 모델의 최적의 초모수를 이용하여, 모델을 최종적으

로 학습하는 과정이다. 해당 모델에 기존 데이터를 입력하게 되면 새로운 사고율에 대한 값을 출력할 수 있다.

```
# 모델 학습
model_life = catboost.CatBoostRegressor()
model_life.fit(
    X_train,
    y_train,
    cat_features=life_cat_col
)
```

(2) 사고율에 영향을 주는 변수 분석

첫 번째 방식으로는 모델링 과정에서 산출되는 변수중요도를 해석할 수 있다. 이는 종속변수를 예측할 때 유의미한 정보를 가지고 있는 변수를 파악하는 과정이다. 파생변수의 효능을 검증하고, 새로운 파생변수를 위해 변수 간 상호작용을 파악할 수 있다. 또한 이 과정을 통해 불필요한 변수를 제거한다.

```
# 변수 중요도를 출력한 후, 변수 간의 상호작용을 파악
model_life.get_feature_importance(prettified=True).head(10)
model_life.get_feature_importance(prettified=True, type='Interaction').head(10)
```

하단의 그림은 해당 코드의 결과이다.

변수 중요도 출력

각 변수 간 중요도를 내림차순으로 출력한 결과이다. 종속변수인 사고율에 예측하는데 유의미한 정보를 담고 있는 변수이다. 이는 사고율을 모델링할 때 중요한 역할을 할 가능성이 높다.

	Feature Id	Importances
0	OCH_WEDD_EXP_SCORE	5.545802
1	매출금액합계_자동차_D	3.688635
2	ESTUD_OCH_SCORE	3.233115
3	매출금액합계_요식_B	3.200948
4	매출건수합계_요식_D	2.760089
5	매출건수합계_자동차_A	2.632486
6	매출금액합계_교육_C	2.568393
7	총사용금액_B	2.522461
8	매출금액합계_자동차_F	2.502392
9	매출금액합계_교육_A	2.466469

변수 중요도를 통한 변수 간 상호작용 파악

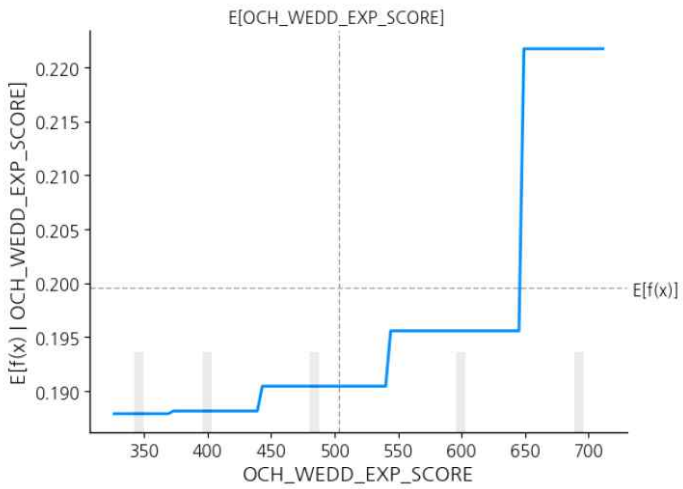
두 변수 간의 상호작용 값을 내림차순으로 출력한 결과이다. 상호작용이 높은 변수끼리 묶어 새롭게 파생변수를 생성하는 방식을 적용할 수 있다.

	First Feature Index	Second Feature Index	Interaction
0	14	61	0.446203
1	68	78	0.436317
2	62	75	0.378489
3	25	58	0.332857
4	21	70	0.327232
5	20	74	0.323630
6	38	83	0.318340
7	45	79	0.317360
8	75	85	0.306171
9	17	79	0.283650

두 번째 방식으로는 Shap 라이브러리를 이용하여 Partial Dependence Plot(PDP) 그래프를 출력하여 사고율에 영향을 끼치는 변수의 영향을 분석하였다. 이는 하나의 독립변수가 최종 사고율 예측에 평균적으로 어떤 영향을 끼쳤는지 수치적으로 보여준다.

```
# Partial Dependence Plot(PDP) 그래프 출력
explainer = shap.Explainer(model_life)
shapely_values = explainer(df_life.drop(['사고율'], axis=1))
shap.plots.partial_dependence(
    'OCH_WEDD_EXP_SCORE',
    model_life.predict,
    df_life.drop(['사고율'], axis=1),
    ice=False,
    model_expected_value=True,
    feature_expected_value=True
)
```

그 출력값은 다음과 같다.



Partial Dependence Plot(PDP)

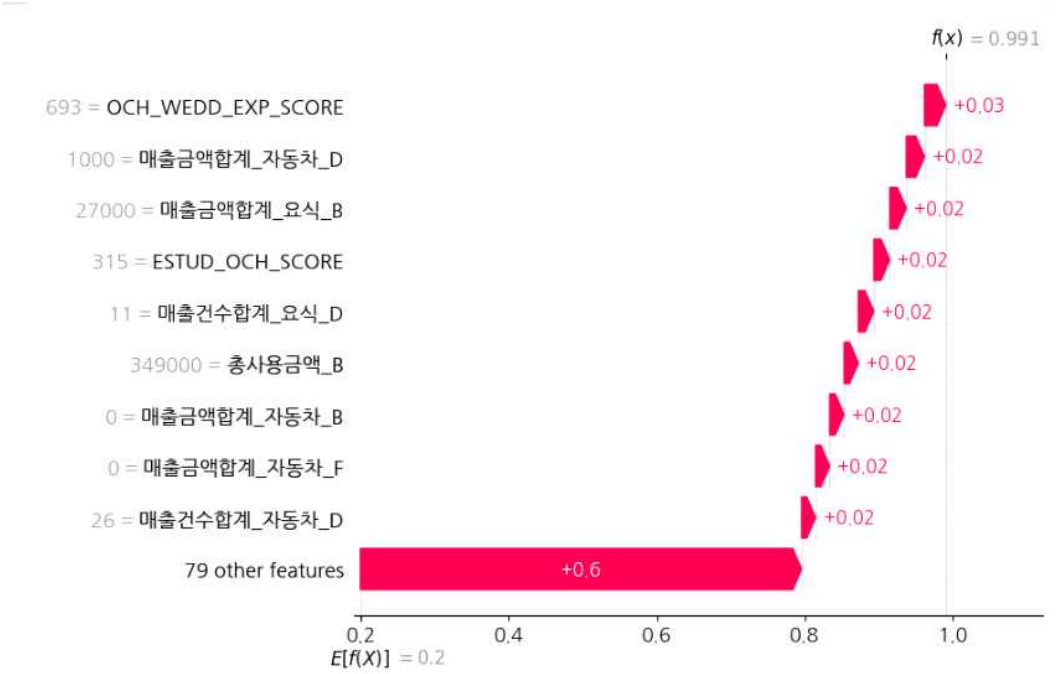
'OCH_WEDD_EXP_SCORE' 라는 독립변수의 값이 커질수록 사고율이 높아진다는 패턴을 확인할 수 있다. 해당 그래프를 통해 독립변수의 값에 따른 사고율에 대한 히스토그램(흰 막대 그림)과 사고율의 기댓값의 변화(파란색 선)를 정확한 수치를 통해 파악할 수 있고, 이는 곧 독립변수 하나의 개별 기여도인 Marginal Contribution을 나타낸다.

세 번째 방식은 마찬가지로 Shap 라이브러리를 이용하여 Shapley value 기반의 영향 분석을 실시할 수 있다. 이때, Shapley value란 특정 하나의 고객(개체)에 대해서 사고율에 대한 각 독립변수의 영향을 파악할 수 있는 지표이다. 즉, 변수의 영향을 해석하는 것뿐만 아니라, 모델의 예측값에 대한 근거를 보여주고, 모델의 신뢰도를 높일 수 있다.

```
# 3번째 고객에 대한 Shapley value 출력
shap.plots.waterfall(shapely_values[2])
```

출력값은 하단의 그림과 같다.

3번째 고객에 대한 Shapley value의 값을 보여주는 그래프로, 종속변수에 영향을 끼치는 각 독립변수의 계수*개체의 실제값의 효과를 파악할 수 있다. 대표적으로 'OCH_WEDD_EXP_SCORE'이라는 변수가 최종 예측값에 +0.03의 영향을 끼쳤다고 해석할 수 있다.



(3) 군집화를 통한 고객 특성 분류 후 이상 고객이 포함된 군집 판별 및 효과성 분석

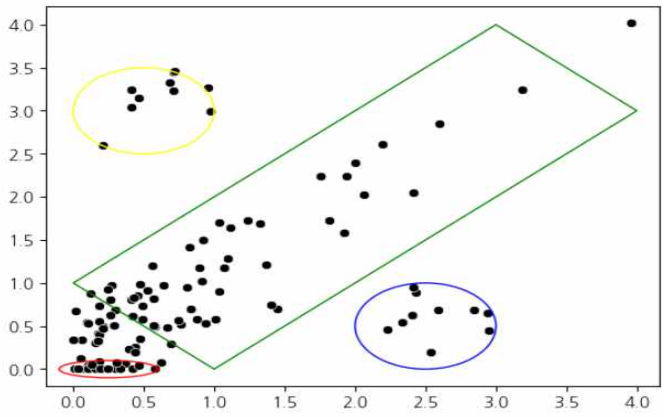
모델의 일반화가 잘 이루어졌고, 카드 명세서 데이터와 보험 명세서 데이터에 대한 패턴을 효과적으로 학습했다는 전제 하에, 고객에 대한 특성을 분류하여 이상치에 속하는 고객을 판별할 것이다. 이때 분류하는 과정에서 있어, 군집화를 하기에는 샘플 데이터만으로는 부족하기 때문에, 각 변수에 대한 난수 생성을 통한 모의 데이터를 이용해 군집화를 진행하였다. 군집화에 적용될 변수는 실제 사고율과 모델링을 통해 출력된 예측 사고율이다. 두 변수를 바탕으로 2차원 상에 이를 표현하여 밀도 기반의 군집화 방식인 (H)DBSCAN으로 군집화를 수행할 것이다. 실제 데이터의 분포가 고르게 분포되어 있지 않음이 예상되기에, 군집의 모양에 제약이 있는 K-Means 방식이 아닌 해당 방식을 사용하는 것이 효과적이다.

```
# 군집화 시뮬레이션을 위한 난수 생성
pred_accident, true_accident = make_regression(n_samples=100, n_features=1, noise=8)
pred_accident, true_accident = abs(pred_accident).flatten(), abs(true_accident).flatten()
data = list(zip(pred_accident, true_accident))
x = np.random.exponential(scale=1.2, size=100) /2
y = x + np.random.normal(size=100) /2
main_trend = list(zip(np.maximum(0, x), np.maximum(0, y)))
anomaly1 = list(zip(np.random.uniform(size=10), np.random.uniform(low=2.5, high=3.5, size=10)))
anomaly2 = list(zip(np.random.uniform(low=2, high=3, size=10), np.random.uniform(size=10)))
data = main_trend + anomaly1 + anomaly2
pred_accident, true_accident = zip(*data)


# 군집화 예상 결과 시각화
ax = sns.scatterplot(x=pred_accident, y=true_accident, color='black')
highlight1 = patches.Ellipse(xy=(0.25, 0), width=0.7, height=0.2, color='red', fill=False)
highlight2 = patches.Polygon(
    xy=[
        [0, 1],
        [1, 0],
        [4, 3],
        [3, 4]
    ],
    color='green',
    fill=False
)
highlight3 = patches.Ellipse(xy=(0.5, 3), width=1, height=1, color='yellow', fill=False)
highlight4 = patches.Ellipse(xy=(2.5, 0.5), width=1, height=1, color='blue', fill=False)
ax.add_patch(highlight1)
ax.add_patch(highlight2)
ax.add_patch(highlight3)
ax.add_patch(highlight4)
```

임의 데이터를 통한 군집화의 결과를 시각화한 그림은 하단과 같다.

```
# 시각화 결과
```



위 그림과 같이 실제 사고율 대비 예측 사고율 그래프를 보면 총 4개의 군집을 관찰할 수 있을 것으로 예상된다. 만약 실제 데이터가 모의 데이터와 같이 구분이 선명하다면 수작업으로 군집화를 진행할 것이고, 그렇지 않다면 (H)DBSCAN이나 Gaussian Mixture 모델을 고려할 것이다.

총 4개의 군집에 대해 각 군집별로 Shapley 값을 기반으로 특정 예측값이 산출된 이유에 대해 분석할 것이다. 특히 예측 사고율이 낮지만, 실제 사고율이 높은 군집은 보

험사에게 가장 손해가 큰 집단, 즉 이상 고객에 대한 집단(군집)이기에 더욱 집중적인 분석을 통해 집단의 예측 오류 원인을 파악해야 한다.

이와 같이, 군집화를 진행하여 고객을 집단별로 분류한 후 이상 고객에 대한 집단을 찾는 과정을 카드 명세서 데이터를 포함하지 않은 본래의 보험 명세서 데이터에 마찬가지로 적용하여 반복하게 된다.

이를 통해 보험 명세서 데이터만을 사용한 사고율과 카드 명세서 데이터를 함께 사용한 사고율을 기반으로 한 각 고객의 군집을 출력할 수 있고, 이상 고객이 포함된 군집 역시 파악할 수 있다. 그 후, 카드 명세서 데이터를 사용한 후 이상 고객이 감소함을 확인함으로써 카드 명세서 데이터의 추가로 인해 사고율을 더 정확히 예측할 수 있다는 가설을 입증할 수 있다.

(4) 이상 고객의 패턴 분석

마지막으로는 사고율에 큰 영향을 끼치는 변수를 파악하는 방식으로 이상 고객에 대한 패턴을 분석하며 보험사가 이상 고객을 대상으로 사고율을 측정할 때 간과한 부분을 살펴볼 것이다. 그리고 해당 변수에 가중치를 주는 방식 등으로 보험료를 올바르게 산출하는 전략을 연구해 볼 것이다. 이때, 변수의 효과 분석 방법 역시 (2)번에 활용한 Shapley value의 값을 사용한다.

ㅇ 개발 도구

구분	도구 이름	버전	제조사(출처)	용도
1	Python	3.10.12	Python Software Foundation	데이터 분석
2	Pandas	1.5.3	PyData	데이터 처리
3	NumPy	1.22.4	NumPy	배열 계산 및 난수 생성
4	Catboost	1.2	Yandex	머신러닝 모델 학습
5	Shap	0.42.0	Scott Lundberg	모델 해석 및 예측 결과 분석
6	Seaborn	0.12.2	PyData	데이터 시각화
7	Matplotlib	3.7.1	The Matplotlib Development Team	데이터 시각화
8	Sklearn	1.2.2	Scikit-Learn	머신러닝 모델 검증

□ 예상 결과

먼저 주기적으로 전산상에 입력되는 고객들의 카드 거래 내역 데이터를 모델이 읽어오고 분석하게 된다. 이때, 고객의 보험 유형에 맞게 총 3개의 모델로 분리하여 사고율을 예측한다. 예측한 사고율은 새로운 보험료 산출에 활용할 수 있다.

고객별 신규 사고율이 예측된다면 기존에 있던 사고율과 비교 검증을 수행해야 한다. 만약 기존 사고율과 새롭게 예측된 사고율 간의 차이가 심하다면, 어떤 원인에 의해 이런 괴리가 발생했는지 면밀히 분석해야 한다. 해당 차이는 분석에 활용되었던 신용카드 거래 데이터의 변수들로부터 파생될 가능성이 크다. 따라서 어떤 부분에 있어서 모델이 그와 같은 결과를 도출해냈는지, 그리고 왜 기존의 사고율과 차이가 많이 발생하게 되었는지 모델의 의사결정 과정, 그리고 Feature importance 등을 자세히 살피며 파악할 수 있게 된다.

본 시스템을 통해 기대할 수 있는 분석 결과 및 효과는 다음과 같다:

1) 고객 소비패턴에 기반한 정확한 사고율 예측

객관적, 합리적이고 적정 보험료를 산출하는 것은 보험사의 수익 관리 및 손실 최소화 관점에서 매우 중요하다. 또한 많은 보험계약자의 이익을 보호함과 동시에 보험자 간의 지나친 경쟁을 억제하여 보험경영을 보호하기 위해서도 필요하다. 그러나 카드 거래 내역 등의 마이데이터를 활용하지 않은 기존 보험료 시스템은 각 고객의 개별 특성을 면밀하게 고려하지 못한다는 점에서 예측력에 한계가 있었다. 단순 개인 인적 사항으로 보험료를 판단하게 되어 수지상등의 원칙에 위배될 확률적인 제약이 분명히 존재한다.

적절한 보험료 산정에 있어 중요한 것은, 고객의 사고율이 높은지 낮은지의 문제가 아니다. 보험사가 비즈니스 과정에서 손실을 입는 큰 원인은 고객들의 예상 사고율에 맞는 보험료를 알맞게 산정하지 못해서다. 즉, 사고율이 현저히 낮은데도 불구하고 사고 발생 시 지급하는 보험료를 너무 높게 측정해버린다는지 등의 이유로 인해 손실이 발생한다. 여러 명의 고객에 대하여 이와 같은 상황이 발생하면 보험사가 입게 되는 손실은 가히 말할 수도 없을 정도로 치명적일 것이다.

하지만 마이데이터와 같은 개인의 특성이 추가된다면 이를 수리적인 알고리즘으로 사고 확률을 제대로 측정하여 수지상등의 원칙에 부합하는 합리적인 방식으로 보험료를 산출할 수 있게 된다. 해당 모델을 통해 고객의 소비 패턴, 이동량, 건강 수준 등을 종합적으로 면밀하게 분석하여 사고율을 정확하게 예측할 수 있다. 이는 궁극적으로 보험료를 적절하게 산정하게끔 하여 보험사의 수익을 극대화하고 손실을 최대화할 수 있게 된다. 즉, "합리적인 보험료"를 산출하게 되는 모델이다. 또한 적정한 보험료의 산출을 통해 보험계약자의 이익을 보호함과 동시에 보험자 간 지나친 경쟁을 억제하며 보험경영을 보호할 수 있게 도와준다.

2) 사고율에 영향을 주는 잠재 요인 파악

사고율에 영향을 주는 주요 요인을 파악하는 과정은 "설명가능한 모델"의 관점에서 고객 신뢰 형성 및 유지에 큰 도움이 된다. 고객이 어떤 패턴을 선보이고, 어떤 요인들이 영향력을 미치는지 모델을 통해 면밀하게 검토하여 맞춤형 보험상품 및 보험료 선정이 가능해진다. 그리고 해당 상품 및 보험금 산정의 기준에 대하여 고객이 의문을 가졌을 경우, 모델의 논리적인 의사결정 과정과 객관적인 수치에 입각하여 설명이 가능하다. 이러한 설득 과정을 통해 자연스럽게 해당 보험사에 대한 신뢰를 높이고, 중도 계약 해지율 역시 감소할 가능성이 높다. 정확한 데이터를 기반으로 산정된 보험료임을 보여준다면 보험 산업에 대한 소비자의 부정적인 인식 역시 변화할 것이다.

또한, 보험사의 입장에서 각 보험의 세부내역별 종류에 따라 사고율에 집중적으로 영향을 끼치는 요인을 찾게 된다면, 해당 요인에 대한 가중치를 주는 방식 등의 집중적 관리로 더 합리적으로 보험료를 산출하고 고객을 관리할 수 있다. 예를 들어 암 발생에 대한 사고율에 가장 많은 영향을 끼치는 변수가 '건강/건강보조업종 매출건수' 라면, 고객의 해당 정보에 대한 더 자세한 정보(과거 이력, 예측값 등)를 조사하고 위험율 예측 모델에 반영하여 더 면밀한 보험료 산출이 가능하다.

3) CRM: 맞춤형 고객 관리 전략

실제 사고율과 예측 사고율의 높고 낮은 정도에 따라 총 4가지의 페르소나(1: 실제 사고율과 예측된 사고율이 모두 낮은 경우, 2: 실제 사고율은 높으나 예측된 사고율은 낮은 경우, 3: 실제 사고율은 낮으나 예측된 사고율은 높은 경우, 4: 실제 사고율과 예측된 사고율이 모두 높은 경우)를 산정할 수 있다. 1&4의 경우 예측 사고율과 실제 사고율 간의 방향성이 똑같기에 예측이 잘 수행된 경우라고 보면 되고, 반대로 2&3의 경우는 예측 사고율과 실제 사고율 간의 괴리가 있기에 주의 깊게 살펴야 할 유형이다. 군집화를 진행하여 분류된 4가지 페르소나들의 비즈니스 함의는 다음과 같다:

[1] 우량 고객: 실제 사고율과 예측 사고율 모두 낮은 고객
보험사의 입장에서는 가장 이상적인 유형의 고객이라고 볼 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" 관점의 전략을 취해야 한다. 프로모션이나 인센티브(ex) 로열티 혜택, 보험 갱신 시 할인)를 지급하고 추가적으로 새로운 보험상품이나 서비스로의 유도 등을 제안해보는 전략을 생각할 수 있다. 해당 서비스에서 이탈하지 않도록 하는 것이 제일 중요하다.

[2] 이상 고객= 집중 관리 대상: 실제 사고율은 높으나 예측 사고율은 낮은 고객
보험사 입장에서 가장 경계를 해야 되는 유형이다. 보험사의 수익 구조에 손실을 가져다주는 위험 고객이기 때문이다. 예측 사고율이 낮아 고객이 지불해야하는 보험료가 필요 대비 적은 수준으로 측정이 될 것이고, 궁극적으로는 보험사 측의 막대한 손실로 이어질 수 있다. 이 경우는 손실을 방지하기 위해 어느 정도 고객 보험료를 높이는 방향으로 나아가야

할 것이다. 그 외 보험료 할증 적용이나 특정 고위험 활동에 대한 보험 보장의 제한을 두는 등의 방식을 통해 사고율을 낮추는 방향으로 가야 한다.

[3] 관심 고객 : 실제 사고율은 낮으나 예측 사고율은 높은 고객

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. 실제 사고율이 낮아 체감상 사고가 적게 발생하는데도 불구하고 과도한 보험료를 지불하는 고객들이 존재한다는 뜻이다. 이는 궁극적으로 장기적인 관점에서 타 보험사로의 서비스 이탈 및 중도 해지 등으로 이어질 수 있다. 따라서 해당 고객들은 보험료를 기본적으로 낮추고, 자신에게 설정된 보험료에 대한 객관적인 수치와 자료를 통한 설득이 필요하다. 이를 통해 고객의 신뢰를 형성하며 보험사는 돈을 벌기 위해서 고객을 끌어들인다는 왜곡된 인식을 해결할 수 있고 정확한 데이터를 기반으로 한 보험료임을 보여줄 수 있다.

[4] 주의 고객: 실제 사고율과 예측 사고율 모두 높은 고객

예측이 잘 이루어지고 있지만, 사고 발생 확률이 높은 고객들이기 때문에 보험사 입장에서는 보험금을 많이 지불하게 될 수도 있다. 이러한 위험을 관리하기 위해서는 궁극적으로 고객들의 사고율을 낮추는 방향으로 유도하는 전략이 좋다. 해당 고객들이 어떤 요인으로 인해 사고율이 높게 측정되는지 앞서 제시한 Feature importance 등의 방법론을 활용하여 살펴본다. 그리고 해당 고객들에게 "고객 안전성"을 강조하며 고객이 왜 사고 발생 위험율이 높게 측정되고 이를 방지하기 위해서는 어떤 조치를 취해야 하는지 설명을 해주는 서비스를 생각할 수 있다. 가령 또한 운전 사고 발생 확률이 높은 고객의 경우 운전강좌를 수강하도록 적극 권유하고, 이런 정기적인 예방 조치를 취하는 고객에게는 일정 수준의 할인 또인 환급을 제공하는 방안을 생각할 수 있다.

이렇게 실제 사고율과 예상 사고율을 바탕으로 고객의 유형을 분류하여 고객에 대한 맞춤형 관리 전략을 세울 수 있다. 각 고객의 유형의 특징을 파악하면서 적합한 고객 관리 방안을 실시하면 된다. 특히, 이상 고객의 유형을 식별하고 그 패턴을 분석하면서 보험사가 간과했던 변수(요인)에 대해 자세히 파악하는 것이 중요하다. 해당 요인들을 사용하여 고객의 실제 사고율 산정에 가중치를 두고 모델의 재학습을 통해 더 현실적인 보험료 산정을 할 수 있게 된다.

4) 신규 가입 고객에 대한 대응력 강화

아무리 좋은 모델일지라도 100% 완벽하게 예측할 수는 없다. 본 시스템도 마찬가지다. 하지만 해당 모델의 차별점은 보험료가 잘못 산출된 경우, 이에 대한 대응을 빠르게 하고 업데이트를 다시 효과적으로 할 수 있다는 점에 있다. 즉, 문제가 발생하여 보험료가 잘못 산출되었을 시 원인을 빠르게 파악하여 보험료를 다시 산출 후 재분석을 할 수 있다. 이는 해당 모델이 Explainable AI 방법론에 입각한다는 점에 기인한다. 손실 분석 과정에 있어 어떤 변수가 보험료 산출에 악영향을 끼쳤는지를 파악할 수 있기 때문이다. 나아가 거시적인 시각에서, 현재 보유한 데이터를 통해 보험사의 수익원, 손실원이 무엇인지를 정확하게 제시할 수 있다는 점에서 의의가 있다.