자동차보험 사고율 예측에 기반한 고객 맞춤 특약 추천 마케팅 제안

목차

1. 분석배경

1.1 전략 재수립을 위한 사고율 예측 모델의 필요성

코로나19가 완화된 이후 자동차의 통행량이 증가함에 따라, 보험사의 자동차보험 손해율은 지속적으로 증가하는 추세를 보이고 있다. 23년 상반기 기준 손해보험사들의 손해율은 78.0%로, 전년 동기(77.1%)에 비해 약 0.9%p 상승한 모습을 보였다.

따라서 보험사는 가입 희망자 및 기존 고객 데이터를 종합적으로 반영할 수 있는 사고율 예측 모델을 개발을 통한 환경 변화에 따른 전략을 재 수립할 필요가 있다. 모델을 통해 얻은 사고율 예측 계산 결과로 보험사는 변화된 부분을 복합적으로 고려해 정교한 손해율 관리를 진행할 수 있을 것이다. 또한 환경의 변화로 인해 손해율이 상승함으로써 발생할 수 있는 문제들을 최대한 예방할 수 있을 것이다.

1.2 특약을 활용한 초개인화 서비스의 필요성

보험업계에서 낮은 사고율과 연관성이 높은 특약을 통해 우량 고객을 모집하려는 이른바 '특약경쟁'이 심화되고 있는 상황이다. 보험사마다 특약의 혜택을 강화하며 우량 고객 모집에 적극적으로 나서고 있다. 그러나, 사고율이 낮은 우량 고객을 유치하기 위해 내걸었던 블랙박스 특약의 경우, 블랙박스 특약 가입자의 사고율은 25.34%로 미가입자 24.2%보다 높았기에 블랙박스 관련 특약의 실효성에 대한 문제가 제기되고 있다.

이러한 문제점에 주목해 각 개인의 특성을 파악해 특약의 실효성을 높일 수 있는 초개인화 전략이 필요하다고 판단했다. 모든 고객에게 일괄적으로 특약 상품을 추천해주는 것이 아닌 보험사 입장에서 우량 고객이 될 수 있는 고객군의 특징을 파악해 해당하는 고객들에게 적합한 특약을 추천해줌으로써 특약의 실효성을 높여 문제를 해결할 수 있을 것이라 생각한다. 따라서, 사고 발생률의 위험요인과 완화요인을 파악하고 각 고객의 특성을 반영한 특약 추천 방안을 제시하고자 한다.

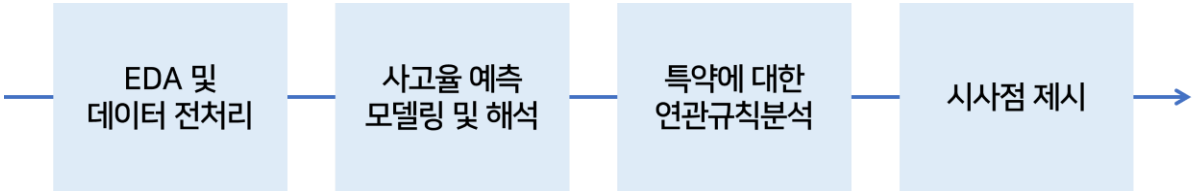
2. 데이터 개요 및 분석 흐름

제공받은 데이터 셋은 2022년 7월부터 2023년 6월까지의 자동차 보험 가입 고객들을 대상으로 한 마감 실적의 요약 정보이다. 분석의 편의성을 위해 모든 변수명은 한글로 변환하여 분석을 진행했으며, 이 중 '운전자한정특별약관' 변수의 경우 '운전자한정특별약관코드'와 동일한 내용을 포함하고 있어 분석에서 수치형으로 표현된 '운전자한정특별약관코드'만을 분석에 사용했다. 또한 '상품대분류코드_자동차' 변수의 경우 모든 관측치가 동일한 값을 갖는 것으로 파악되어 해당 변수를 제거하고 분석을 진행했다. 해당 과정을 거쳐 데이터가 11개의 설명변수와 목적변수의 계산에 사용되는 2개의 변수(유효대수, 사고건수)로 구성되어 있음을 확인했으며, 각 변수명에 대한 설명은 아래와 같다.

변수명	정의
피보험자연령대	자동차 계약 보험 증권에 피보험자로 기재되어 있는 기명 피보험자의 연령대
피보험자성별코드	001:여성, 002:남성
국산구분코드	1: 국산, 2:외산
직전3년간사고건수	N:무사고, D:3년1회, C:3년2회, B:3년3회, ZZZ:기타(피보험자로 해당차량 가입 3개월이내 신규계약)
차량경과년수	신차, 5년이하, 10년이하, 10년초과
차종	소형A:1000cc이하, 소형B:1600cc이하, 중형:2000cc이하, 대형:2000cc초과 , 다목적(1종,2종) 승차정원 7인이상 10인이하
운전자한정특별약관	1: 누구나(기본), 2: 가족한정(형제자매제외), 3: 부부한정, 4: 기명피보험자1인한정, 5: 가족및형제자매한정, 6: 가족및지정1인, 7: 1인및지정1인, 10: 부부및지정1인, 11: 임직원한정, 12: 부부및자녀한정
가입경력코드	01:1년미만, 02:2년미만, 03:3년미만, 04:4년미만, 05:5년미만, 06:6년미만, 07:7년미만, 08:7년이상
차량가입금액	자차미가입, 5천만원이하, 1억이하, 1억초과
영상기록장치특약요율	0:블랙박스 미가입, 1:가입
마일리지약정거리	마일리지 약정별 예상 할인율
유효대수	경과일수에 따른 가입대수
사고건수	유효대수 기간 중에 발생한 사고건수

(표1) 데이터 레이아웃

우선, 제공받은 데이터셋을 활용해 본 분석의 목적변수가 되는 사고율을 계산했다. 이 때, 사고율 계산에 사용되는 '유효대수'는 기존 산정 방식에 의거해 반올림하여 소수점 이하의 단위는 반영하지 않도록 조치했으며 '사고건수 ÷ 유효대수' 로 사고율을 계산했다. 이 때, 분모에 해당하는 유효대수 변수가 0인 경우는 사고율 계산이 되지 않아 데이터에 노이즈가 있는 것으로 판단해 제거했다. 이를 통해 99,893개의 관측값이 제거되어 167,884개의 관측값과 14개의 변수로 이뤄진 데이터셋을 구성했으며, 다음과 같은 흐름으로 분석을 진행하고자 한다.



(그림1) 분석 흐름 도식

3. EDA 및 데이터 전처리

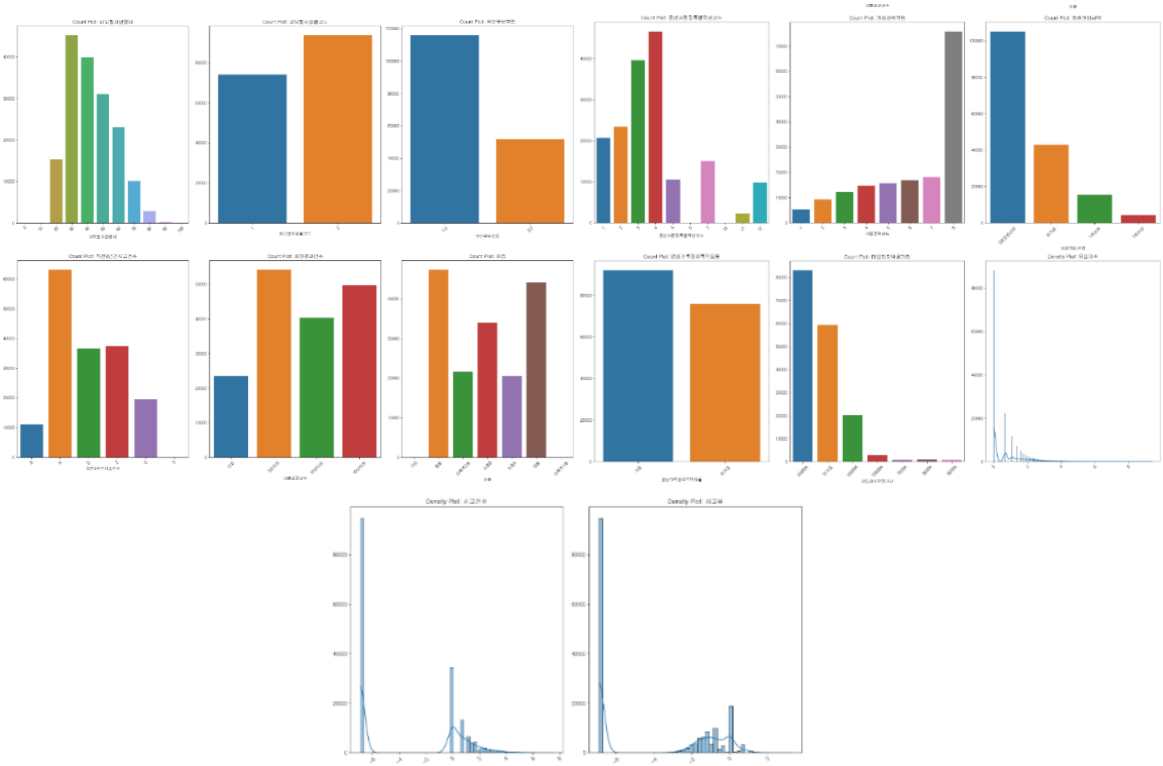
3.1 EDA

3.1.1 데이터 기본 정보

각 변수의 자료형 및 고유값 수를 확인한 결과, '유효대수', '사고건수', '사고율' 등의 변수는 실수형이며 고유값이 많은 것으로 보아 연속형 데이터임을 확인했다. 이외의 변수들은 고유값의 수가 2~10개로 적고 데이터 레이아웃의 정의를 확인한 결과 범주형 데이터임을 확인할 수 있었다.

3.1.2 변수별 분포 시각화

각 변수별 분포를 시각화한 결과는 다음과 같다. 이 때, 연속형 변수의 경우 0인 값과 이상치로 인해 시각화를 통한 분포의 확인이 어려워 로그 변환을 통해 스케일링을 진행한 후의 값을 시각화했다.



(그림2) 변수별 분포 시각화

우선 연령대의 경우, 30대가 가장 많았으며 10대 이하와 90대 이상이 매우 적은 값을 보였다. 해당 변수의 경우, 운전자의 연령대를 나타내는 것임에 반해 0~10대와 90대 이상은 이상치로 판단될 여지가 있다고 생각하여 본 분석에서는 제외하는 것으로 결정했다. 이를 통해, 253개의 관측값이 제거되었다.

가입자의 성별은 남성이 여성에 비해 약 1.3배 많았으며, 국산차를 운전하는 운전자가 외산차를 운전하는 운전자에 비해 약 2배 가량 많은 것으로 나타났다. 직전 3년간 사고건수를 확인한 결과, 무사고인 운전자의 비율이 가장 높았으며 신규가입자, 사고1회, 사고2회, 사고3회의 순서로 나타나는 것을 확인할 수 있었다.

차량 경과년수에 따른 가입자의 비율은 5년 이하의 차를 운전하는 운전자의 비율이 가장 높았으며 신차의 운전자 비율이 가장 적었다. 차종의 경우 중형차와 대형차 운전자가 가장 많았으며, 다목적 1종 차량 및 기타 차량 운전자의 수가 매우 적게 집계된 것을 확인할 수 있었다.

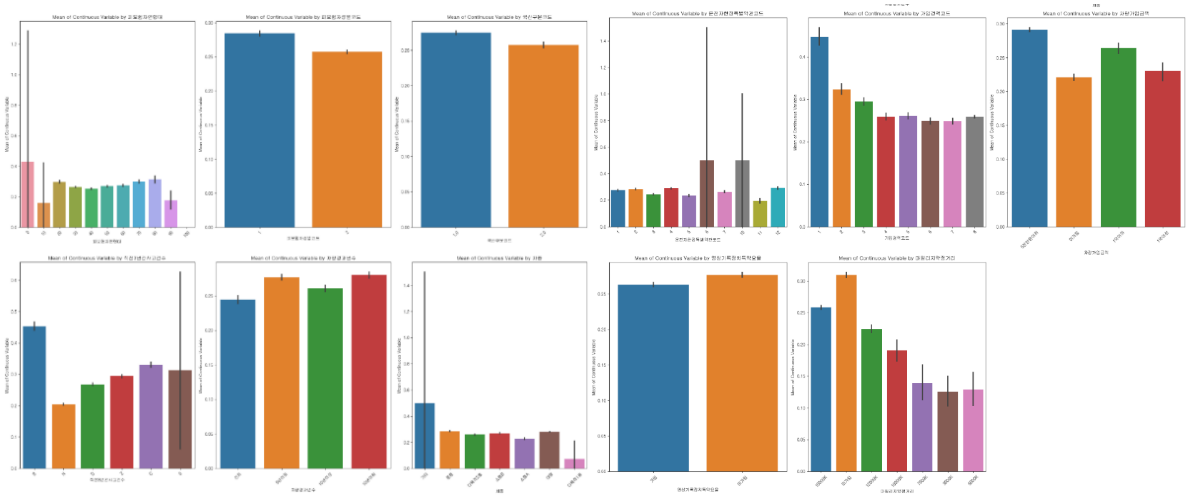
운전자한정특약에서는 '기명피보험자1인한정(4)'의 운전자가 가장 많아 1인 보험의 가입자가 가장 많은 것을 추론할 수 있었으나, '가족 및 지정1인(6)'과 '부부 및 지정1인(10)' 및 '임직원한정(11)'은 그 집계값이 매우 낮게 나타난 것을 확인할 수 있었다.

가입경력에 따른 가입자 수는 1년 미만이 가장 적고 8년 이상이 가장 많게 나타나는 것을 통해 경력이 높은 가입자들이 많다는 것을 확인할 수 있었다. 차량가입금액에 따른 가입자 수는 5천만원 이하의 가입자 수가 전체 가입자의 절반 이상을 차지할 정도로 가장 많았음을 확인할 수 있었다.

또한 영상기록장치요율의 가입자가 미가입자에 비해 약 1.2배 많았으며 마일리지 약정거리를 살펴본 결과 15000K의 마일리지 특약 가입자의 비율이 가장 높았으며 그 다음으로 미가입자의 비율이 높았다. 3000K~10000K의 마일리지 특약 가입자는 수가 매우 적은 것을 확인할 수 있었다.

연속형 변수의 경우, 유효대수와 사고건수는 최대값이 12,045와 1,847로 이상치의 영향으로 분포가 왜곡되는 것을 시각화를 통해 확인할 수 있었다. 사고율의 경우 0과 1 사이에 대부분의 값이 관측되어 나타나며, 최대값은 27로 이상치의 영향으로 인해 분포의 불균형이 심한 것을 확인할 수 있었다.

3.1.3 각 변수와 사고율 관계 파악



(그림3) 변수별 범주에 따른 평균 사고율 시각화

각 변수와 사고율 간의 관계성을 파악한 결과, 연령대별 사고율에서 80대가 가장 높은 평균 사고율(0.31)을 보였으며, 40대가 가장 낮은 평균 사고율(0.25)을 보였다.

직전 3년간 사고건수는 무사고 집단의 사고율(0.20)이 가장 낮고 3회 이상 집단의 사고율(0.45)이 가장 높게 나타나 과거의 기록과 현재의 사고율 간에 높은 관련성을 보이는 것을 추론할 수 있었다.

차종에서는 중형차(0.28)의 사고율이 가장 높았으며, 다목적1종차량(0.08)의 평균 사고율이 가장 낮

은 모습을 보였다. 운전자한정 특약에 따른 사고율은 '가족 및 지정1인(6)'과 '부부 및 지정1인(10)'에서 가장 높았으며(0.50), '임직원한정(11)'에서 가장 낮은 모습을 보였다(0.20).

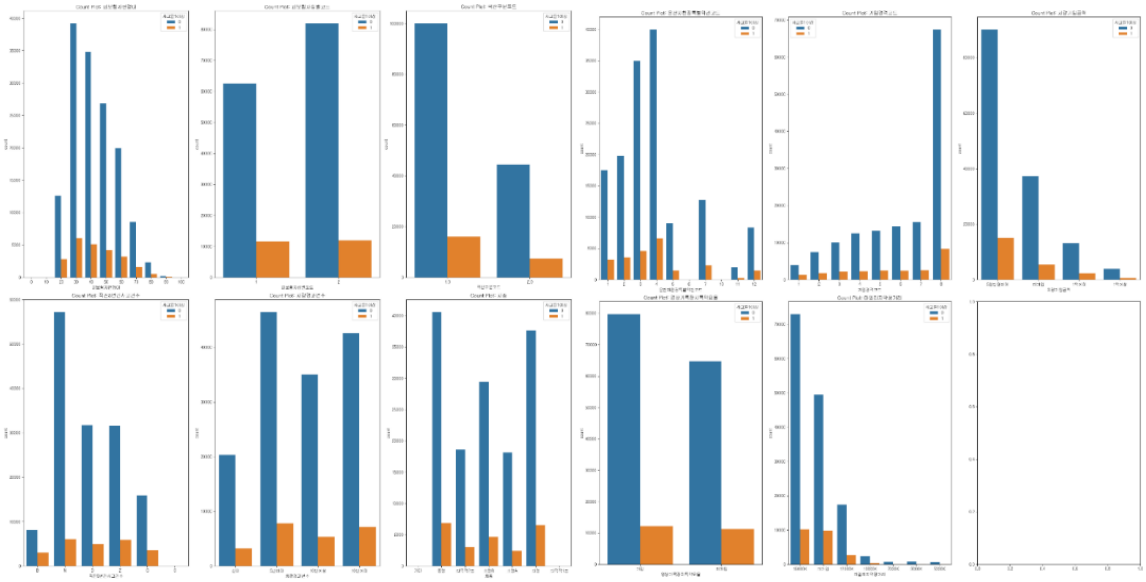
가입 경력에 따른 사고율은 1년 미만 집단의 평균 사고율이 가장 높고(0.45) 경력 7년 집단의 사고율이 가장 낮아(0.25) 경력과 사고율 간 음의 상관관계를 보이는 것으로 추정할 수 있었다. 차량가입금액에 따르면, 미가입 집단의 사고율이 오히려 가장 낮았으며(0.22) 5천만원 이하의 가입 금액을 보이는 집단의 평균 사고율이 가장 높았다(0.29).

마일리지 약정 거리에 따른 사고율에서는 가입자가 가장 적은 3000K~7000K 구간이 대체적으로 낮은 사고율(0.12~0.13)을 보인 반면 미가입자가 가장 높은 사고율(0.31)을 보이고 15000K의 마일리지 가입자 수가 많은 운전자(0.26)가 그 뒤를 잇는 모습을 보였다.

또한 위의 변수에 따른 범주별 평균 사고율 차이의 유의성은 모두 ANOVA[1] 검정을 통해 유의수준 5%에서 그 유의성을 확인했다. 다만, 운전자한정특약, 차종과 같은 일부 변수의 경우, 사고율이 높게 나타난 범주와 낮게 나타난 범주의 관측값 자체가 적어 이상치의 영향으로 인해 그 영향력이 본래의 영향력에 비해 과대 및 과소평가되었을 여지가 존재함을 밝힌다.

3.1.4 변수별 고위험군 분포 확인

한편, 사고율과 기타 설명 변수의 관계를 살펴보던 중, 사고율 1을 기준으로 사고율이 1 이상인 고객과 그렇지 않은 고객 간에 각 설명변수의 분포에서 차이를 보이는 것으로 드러났다. 사고율 1을 기준으로 비교했을 때의 각 변수별 분포는 아래와 같다.



[1] 두 개 이상 다수의 집단의 동질성을 검정하는데 사용하는 방법. 정규성 가정을 필요로 하며, 본 분석에서는 이를 만족시키기 위해 연속형 변수에 로그 변환을 실시하여 정규성 가정을 충족시킨 후 검정을 실시했다.

(그림4) 고위험군 여부에 따른 변수별 분포 시각화

사고율 1 이상을 고위험군으로 분류했을 때 성별, 직전 3년간 사고건수, 가입경력 등과 같이 일부 변수에서 각 범주에 따른 고위험군의 수가 유사하게 나타나는 변수가 많은 것을 확인할 수 있다.

또한 ANOVA 검정 결과, 이같은 고위험군에서는 성별과 보유 차량의 국산외산여부에 따른 사고율의 차이가 유의수준 5% 하에서 통계적으로 유의하게 나타나지 않는 것을 확인할 수 있었다. 이외의 변수에서는 변수별 사고율이 가장 높았던 집단과 낮았던 집단이 동일하게 나타났다.

3.2 결측치 확인 및 대체

각 변수별 결측치를 확인한 결과, '국산구분코드' 변수에 11개의 결측치가 존재하는 것을 확인할 수 있었다. 이에 대해 전체 약 17만개의 데이터에 대해 결측치가 존재하는 11개의 관측치의 영향력이 미미할 것으로 판단해 본 분석 과정에서는 이를 제외했다.

3.3 Encoding (자료형 변환)

'직전3년간사고건수', '차량경과년수', '차종', '차량가입금액', '영상기록장치특약요율', '마일리지약정거리'와 같은 변수들은 자료형이 문자형으로 되어있는 것을 확인했다. 분석을 위해서는 수치형 자료로 변환할 필요가 있어 각 변수의 특성에 따른 적절한 Encoding 기법을 통해 해당 변수의 자료형을 변환하는 과정을 거쳤다.

우선 '차량경과년수', '차종', '차량가입금액', '마일리지약정거리' 변수의 경우 각 범주 간의 위계가 존재하기에 순서형 자료로 파악을 하고, 이에 대한 Label Encoding[2]을 통해 자료를 수치형으로 변환했다. 해당 작업은 '미가입' 변수를 비롯한 가장 작은 크기의 변수를 0으로 표현하고, 그 크기의 증가에 따라 1씩 증가하는 정수형 변수로 표시했다.

한편 '영상기록장치특약요율' 변수의 경우, 영상기록장치에 대한 특약의 가입 여부에 관한 변수이기에 가입자를 0, 미가입자를 1로 표시하는 One-hot Encoding[3]을 실시했다.

또한 위와 같은 변수 이외에, '운전자한정특별약관'의 경우 범주형 자료가 수치형으로 표현되어 있으나 각 범주 간 위계가 존재하지 않는 변수이기 때문에 One-hot Encoding을 통해 더미변수화 하여 자료를 표현했다.

한편, '직전3년간사고건수' 변수의 경우 '기타'와 같은 범주는 그 의미를 명확하게 파악하기 어렵다는 점 등으로 인해 순서형 변수로 해석하기 힘들어 명목형 변수로 간주하여 향후 그 영향력을 상세하

---

[2] Label Encoding : 문자열 자료 혹은 명목형 자료의 각 수준에 임의의 숫자를 할당하는 방법

[3] One-hot Encoding : 데이터에 가변수를 추가하여 인코딩을 진행하는 방법으로, 범주의 수만큼 열을 추가해 해당 범주에는 1, 아닌 값에는 0을 부여.

게 파악하고자 One-hot Encoding을 이용해 더미변수화 하여 자료를 표현했다.

3.4 파생변수 생성

위의 분석 결과와 더불어, 데이터의 정보량을 극대화하기 위해 기존의 변수 정보를 바탕으로 다양한 파생변수를 생성해 분석에 추가적으로 활용했다. 생성한 파생변수는 다음과 같다.

- 직전3년간사고건수_유무 : 직전 3년간의 사고 발생 여부 자체가 사고율에 영향을 미칠 것으로 판단해, 무사고(N)와 신규 가입자(Z)를 사고가 발생하지 않은 것(0)으로, 이를 제외한 나머지 변수를 사고가 발생한 것으로(1) 간주하는 binary 변수 '직전3년간사고건수_유무'를 생성했다.

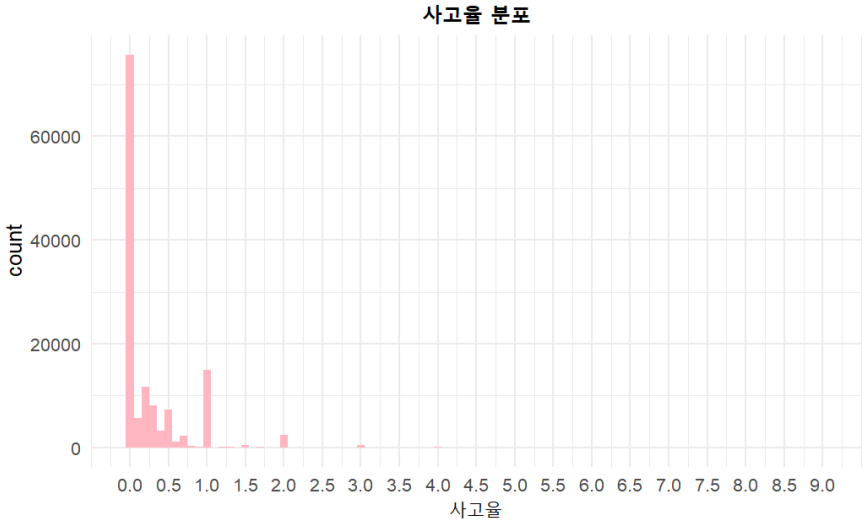
- 마일리지할인율_가입여부 : 마일리지 약정에 따른 할인율 이외에 해당 약정에 대한 가입 여부 자체가 사고율에 영향을 미칠 수 있을 것으로 판단해 '마일리지할인율'의 값이 0인 경우는 0(미가입)으로, 0 초과인 변수의 경우 가입(1)으로 표시하는 binary 변수 '마일리지할인율_가입여부'를 생성했다.

- 자차보험가입여부 : 보험 가입 금액에 관한 변수인 '차량가입금액' 변수를 활용해, 가입 금액 이외에도 보험에 대한 가입 여부 자체가 사고율에 영향을 미칠 수 있을 것으로 판단해 변수 값이 '미가입'인 경우에는 0, 그렇지 않은 경우에는 1(가입)으로 표기하는 binary 변수 '자차보험가입여부'를 생성했다.

- 고경력운전자 및 저경력운전자 : 운전 경력이 사고율에 영향을 미칠 수 있는 요소로 판단해, '가입경력코드' 변수를 기준으로 운전 경력이 8년 이상의 고경력 운전자를 1로 나타내는 이진 변수 '고경력운전자' 변수를, 운전 경력이 1년 미만의 상대적으로 적은 경력을 보유한 운전자를 1로 나타내는 '저경력운전자' 변수를 추가로 생성했다.

4. 사고율 예측 모델링 및 해석

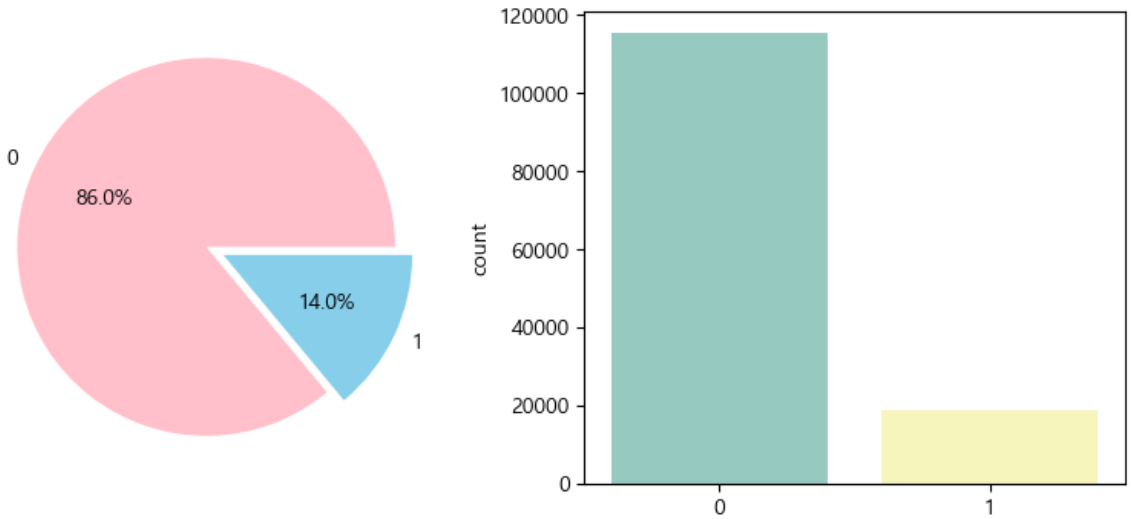
EDA에서 높은 사고율과 관련된 변수가 존재한다는 인사이트를 반영해 데이터 내 패턴을 모델링을 통해 파악하고 이를 이용해 사고율을 예측하고자 했다.



(그림5) Train set 사고율 분포

모델링을 위해 제공받은 데이터의 80%는 train set으로, 20%는 test set으로 분리하여 사용했다. 사고율이 1 이상인 고객을 고위험군으로, 사고율이 1 미만인 고객은 저위험군으로 정의하여 고객군을 분리하여 모델링 진행했다. 고객군을 분리한 이유는 사고율의 분포가 0과 1 사이로 치우쳐져 있어 고객군을 분리하지 않고 모델을 학습시키면 예측 값이 0과 1 사이의 값으로만 도출되어 높은 사고율을 가진 고객을 예측하지 못하는 문제를 해결하기 위함이다. 따라서 고위험군, 저위험군 고객을 로지스틱 회귀 모델로 예측한 후 고위험군 고객과 저위험군 고객에 대한 앙상블 트리 회귀 모델링을 따로 진행해 사고율을 예측했다.

4.1 고위험군/저위험군 고객 이진 분류 모델링



(그림6) 고위험군/저위험군 클래스 분포

train set과 test set에 사고율이 1 이상인 고객을 1, 사고율이 1 미만인 고객을 0으로 고위험군 고객 여부를 나타내는 변수를 추가적으로 생성해 로지스틱 회귀 모델 종속변수로 사용했다. Train set의 클

래스 분포를 확인해본 결과, 저위험군 고객은 train set의 86%, 고위험군 고객은 train set의 14%로 클래스 불균형이 발생했다. 클래스 불균형을 해결하기 위해 로지스틱 회귀 모델의 cut-off point를 0.5가 아닌 FPR과 TPR을 고려한 최적의 cut-off point를 사용하였다. FPR은 실제 음성(0인 클래스) 중 잘못 분류한 비율을, TPR은 실제 양성(1인 클래스) 중 정확히 분류한 비율을 의미한다. cut-off point의 값에 따라 FPR과 TPR의 변화를 보여주는 ROC 곡선에서 FPR이 0에 가까워지고 TPR이 1에 가까워져 분류 예측 시 성능이 가장 정확해지는 곳을 로지스틱 회귀 모델의 cut-off point로 사용했다. 평가 지표로는 정확도가 아닌 F1-Score와 Macro F1-Score를 사용했다. 정확도는 클래스 불균형 상황에서 다수의 클래스에 의존하기 때문에 소수의 클래스 예측 성능에 대해 잘 설명하지 못한다. 따라서, 양성으로 예측한 것 중 실제 양성인 비율인 PPV와 TPR의 조화 평균을 이용해 소수의 클래스를 얼마나 잘 찾는지도 고려할 수 있는 F1-Score과 각 클래스별로 F1-Score의 평균을 내 동등하게 클래스를 고려하는 Macro F1-Score를 평가 지표로 이용했다.

변수명	계수	변수명	계수
피보험자연령대	0.088073	직전3년간 무사고	-0.480064
피보험자성별	-0.214471	직전3년간 사고1건	-0.026261
국산구분	0.164057	직전3년간 사고2건	0.439593
가입경력	-0.101439	직전3년간 사고3건 이상	1.005029
차량가입금액	0.083840	운전자한정특별약관 부부한정 가입	-0.430665
영상기록장치특약요율	-0.123890	운전자한정특별약관 기명피보험자1인한정 가입	-0.245725
마일리지할인특약 가입여부	-0.318825	운전자한정특별약관 1인및지정1인 가입	0.071802
차종	-0.017978	운전자한정특별약관 임직원한정 가입	0.180686
고경력운전자	-0.425987	운전자한정특별약관 부부및자녀한정 가입	0.098966

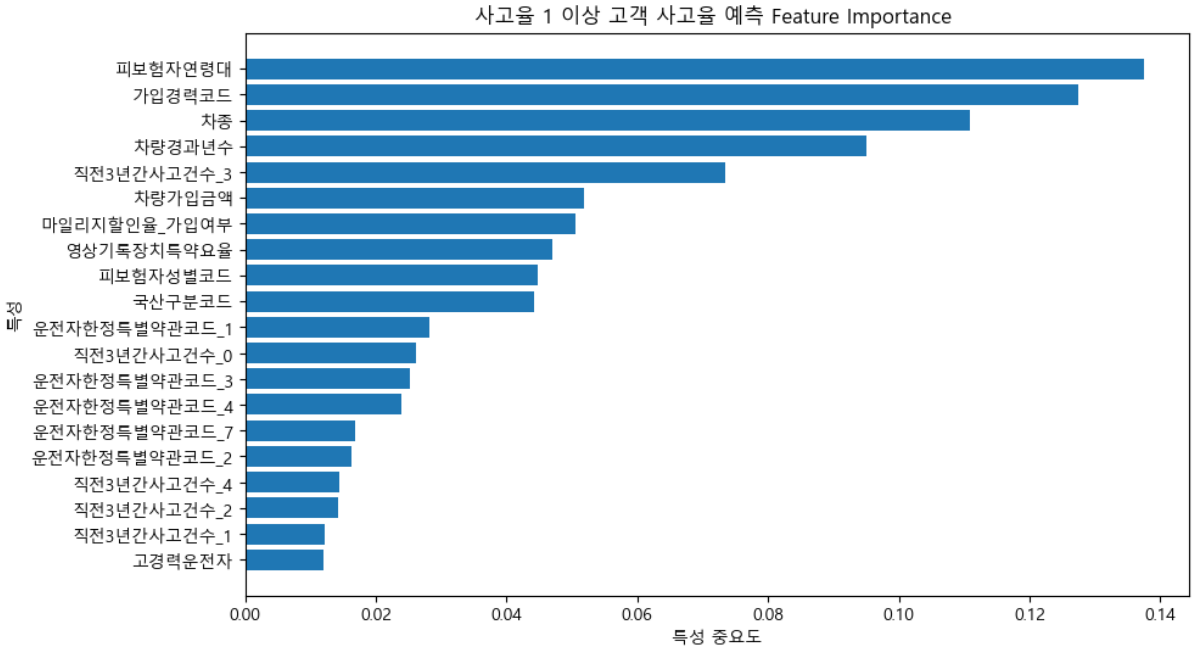
(표2) 로지스틱 회귀 모델 결과

전체 변수들 중 로지스틱 회귀 모델 결과에서 p-value가 0.05 이하로 고위험군 고객과 저위험군 고객을 분류하는데 유의미하다고 나온 변수들로 로지스틱 회귀 모델링을 진행하였다. 피보험자의 가입경력, 직전 3년간 사고건수, 마일리지 할인특약 가입여부가 고위험군 고객을 예측하는데 높은 영향력을 가졌다는 것을 확인할 수 있다. 로지스틱 회귀 모델의 결과를 통해 피보험자의 연령대가 높을수록, 직전 3년간 사고건수가 많을수록 사고율이 높아진다는 것을 알 수 있다. 반면, 경력이 높을수록, 영상기록장치특약과 마일리지 할인특약에 가입했다면 사고율이 낮아진다는 것을 알 수 있다. 해당 로지스틱 회귀 모델의 AUC는 0.66이었으며 cut-off point는 0.1393으로 설정해 이진 분류를 진행했다.

4.2 사고율 예측 모델링

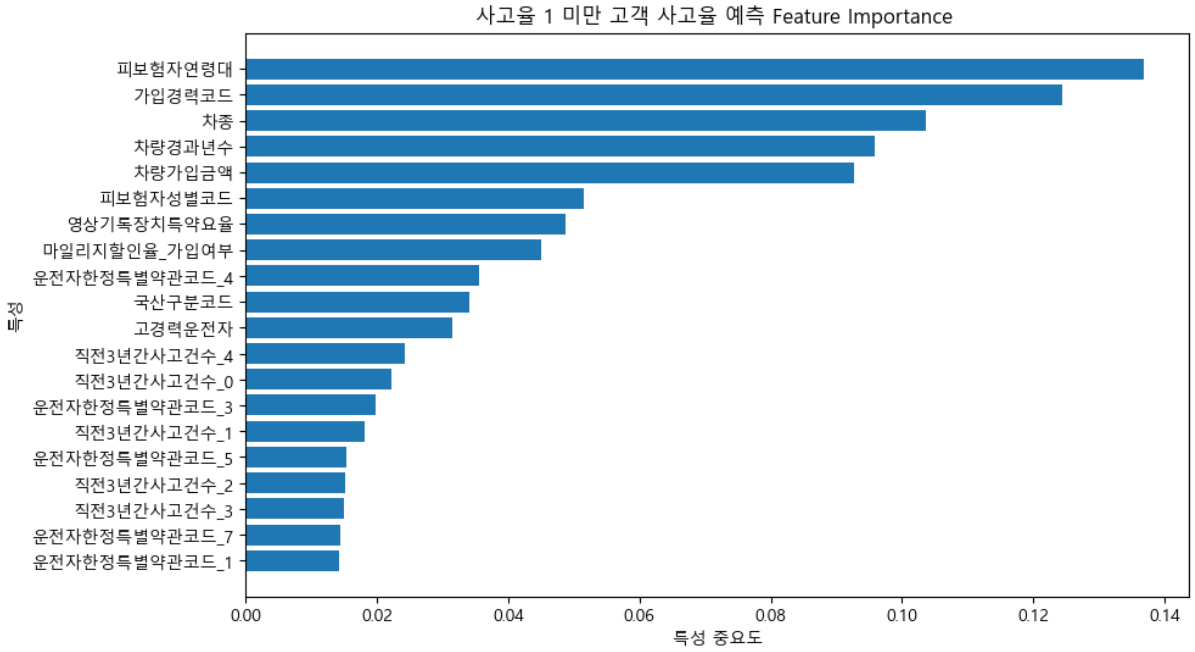
train set에서 고위험군 고객과 저위험군 고객을 따로 분리해 다른 2개의 사고율 예측 모델을 만들었다. 사고율 예측 모델링은 voting을 이용한 앙상블 트리 모델을 이용했다. GLM과 같은 통계 모델은 분포 가정을 필요로 하는 모수적 모델이지만 트리 모델은 분포 가정이 필요하지 않은 비모수적 모델이다. Train set의 사고율 분포의 EDA 결과와 정규분포임을 확인하는 Shapiro-Wilk 검정 결과[4]를 참고했을 때 사고율 분포가 정규분포나 감마분포를 만족한다고 가정하기 어렵다고 판단했다. 따라서 분포 가정을 필요로 하지 않는 트리 모델을 사용했다.

앙상블 모델은 서로 다른 종류의 모델을 결합해 성능을 높이는 모델링 방법으로 각 모델의 예측 값의 평균을 내 최종 예측 값을 도출하는 voting 앙상블 모델을 사용했다. 여러 모델을 결합해 개별 모델을 사용했을 때보다 예측 오차를 줄여 예측 성능을 높일 수 있고 비교적으로 안정된 예측 값을 얻을 수 있기 때문에 voting을 이용한 앙상블 모델이 사고율 예측에 적절할 것으로 판단했다. 다양한 트리 기반 모델을 결합해보면서 가장 성능이 좋았던 Random Forest, Gradient Boosting, Decision-tree regressor 모델을 결합해서 앙상블 모델로 사용했다.



(그림7) 고위험군 고객을 대상으로 한 앙상블 모델 Feature Importances

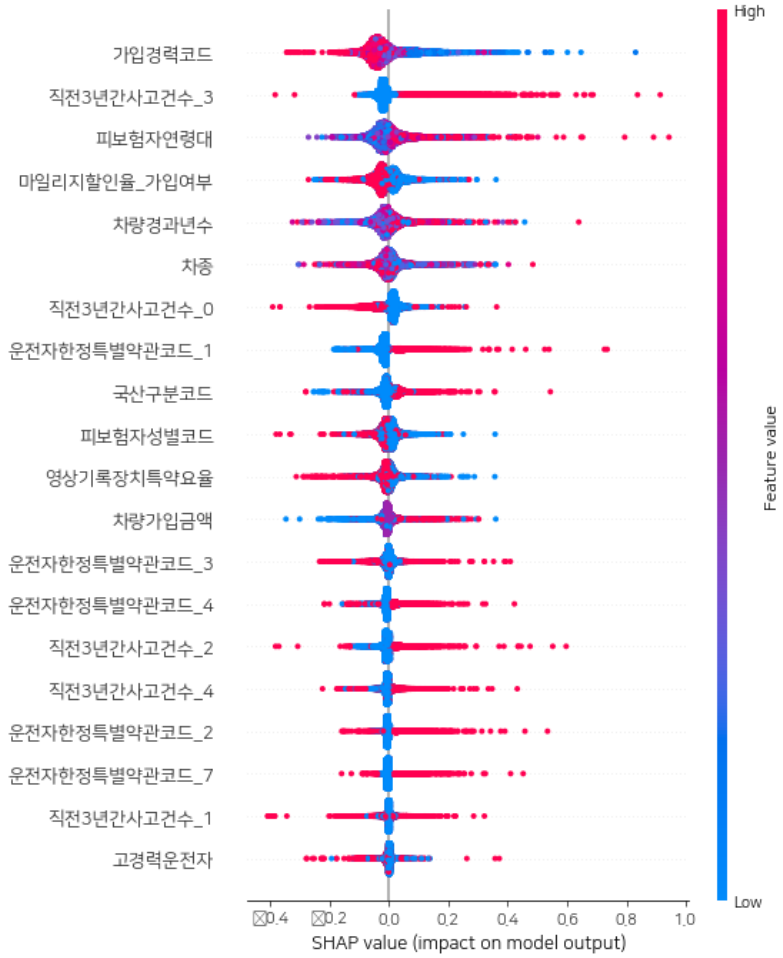
[4] 검정통계량이 0.623731, p-value가 2.2e-16 이하로 데이터가 정규분포를 따른다는 귀무가설을 기각함으로써 정규분포를 따르지 않는다고 판단했다.



(그림8) 저위험군 고객을 대상으로 한 앙상블 모델 Feature Importances

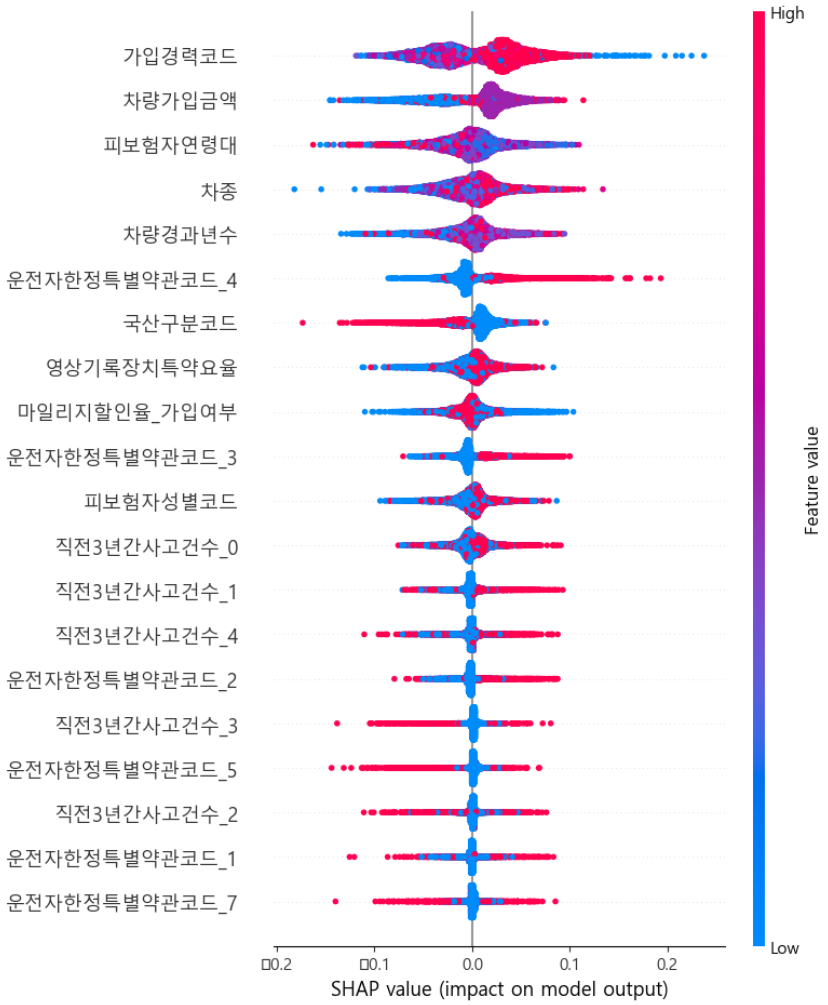
Train set으로 모델을 학습시킨 후 사고율 예측 결과에 대한 각 요인들의 영향력을 파악하기 위해 Feature Importances와 SHAP을 확인했다. Feature importance는 해당 모델에서 각 변수가 예측에 얼마나 중요한 역할을 하는지를 의미한다. SHAP을 통해 변수의 중요도를 평균 낸 값을 활용해 모델에서 얻은 각 예측 값들에 대한 변수의 영향력을 측정할 수 있고 종속변수와 설명변수 간 상관관계를 파악할 수 있다.

두 모델의 Feature Importance를 통해 사고율을 예측하는데 피보험자연령대, 가입경력, 차종, 차량경과년수, 할인특약 여부가 공통적으로 예측에 중요한 역할을 한다는 것을 확인할 수 있었다. 고위험군 고객은 직전 3년간 사고건수가 저위험군인 고객보다 상대적으로 예측에 중요한 역할을 한다는 점에서 차이를 보였다.



(그림9) 고위험군 고객을 대상으로 한 앙상블 모델 SHAP summary plot

고위험군 고객을 대상으로 했을 때 사고율과 뚜렷한 상관관계를 보이는 변수들 위주로 보면 가입경력이 높을수록, 연령대가 높을수록, 직전 3년간 사고건수가 많을수록, 운전자한정특별약관에 가입했다면 사고율이 높아진다는 것을 확인할 수 있다. 반면, 마일리지할인특약과 영상기록장치특약을 가입했다면 사고율이 낮아지는 것을 확인할 수 있다.



(그림10) 저위험군 고객을 대상으로 한 앙상블 모델 SHAP summary plot

이번에는 저위험군 고객을 대상으로 사고율과 뚜렷한 상관관계를 보이는 변수들 위주로 살펴보았다. 운전자한정특별약관 기명피보험자1인한정 및 부부한정을 가입한 경우, 직전 3년간 사고건수가 1건인 경우에 사고율이 높아지는 것을 확인할 수 있다. 반면, 피보험자의 차량이 국산차일 경우에 사고율이 낮아지는 것을 확인할 수 있다. 저위험군의 경우 고위험군의 비해 비교적으로 사고율과 관련된 뚜렷한 상관관계가 없다는 것을 파악했다.

4.3 최종 예측 결과

피보험자 연령대	피보험자 성별	직전3년간 사고건수	…	실제 사고율	예측 사고율
30	여자	무사고	…	1	1.223
40	여자	1건	…	0	0.111
50	여자	무사고	…	0	1.060
20	남자	기타	…	0	1.134

70	여자	2건	...	2	2.258

(표3) 사고율 예측 결과 일부

앞서 진행한 고위험군 고객을 분류하는 로지스틱 회귀 모델을 이용해 test set의 고객들을 고위험군으로 예상되는 고객들과 저위험군으로 예상되는 고객들을 1차적으로 분리했다. 고위험군으로 예상되는 고객들은 train set에서 고위험군이었던 고객들의 데이터로 학습시킨 앙상블 모델에, 저위험군으로 예상되는 고객들은 train set에서 저위험군이었던 고객들의 데이터로 학습시킨 앙상블 모델을 이용해 사고율을 예측했다.

먼저 이진분류모델로 test set의 고위험군 고객과 저위험군 고객을 예측한 결과, F1-Score는 0.3127, Macro F1-Score 0.5232로 도출되었다. 최종적으로 각 앙상블 모델로 test set의 사고율을 예측했을 때 MSE가 0.61로 도출되었다.

고위험군 고객과 저위험군 고객에 대한 분류가 잘 이루어졌다고 가정하고 각 앙상블 모델을 이용해 사고율을 예측했을 때 MSE는 0.076으로 앙상블 트리 모델 자체의 성능은 낮지 않은 편이다. 최종 MSE가 0.61로 비교적 정확성이 떨어지는 이유는 고위험군 고객을 예측하는 이진 분류 과정에서 저위험군 고객을 고위험군으로 실제보다 더 많이 예측해 실제로는 사고율이 낮은 고객이 사고율이 높다고 예측했기 때문이다.

로지스틱 회귀 결과에서 예측에 영향력이 높은 변수로 피보험자의 가입경력, 직전 3년간 사고건수가 뽑힌 것을 통해 운전자의 운전 습관을 직접적으로 파악할 수 있는 변수들이 예측에 영향력을 크게 미친다는 것을 알 수 있다. 이를 통해 교통법규위반 여부처럼 세부적으로 운전자의 운전 습관을 파악할 수 있는 변수들을 추가적으로 활용한다면 예측 모델링에서 더 정확한 사고율 예측이 가능할 것이라 생각한다.

5. 특약에 대한 연관분석(Association Analysis)

5.1 연관분석 활용 배경

SHAP을 통해 사고율과의 상관관계를 확인했을 때, 두 변수 모두 특약에 가입한 경우 사고율이 낮아지는 음의 상관관계를 보였다. 이를 통해 고객이 특약에 가입하도록 유도했을 때 사고율이 낮아질 것이라고 판단하였다. 그러나, 특약에 관심이 없는 고객에까지 가입을 유도하는 마케팅은 비효율적일 뿐 아니라 오히려 사고율이 높은 고객들이 가입해 보험사의 손해율이 높아지는 문제가 발생하기도 했다. 따라서, 특약에 가입할 의사가 높은 동시에 실질적으로 사고율이 유의미하게 감소하는 고객들을 대상으로 추천하는 것이 효율적이라고 판단했다. 연관분석을 통해 특약에 가입할 의사가 높은 고객의 특징을 파악하고 사고율이 유의미하게 감소하는지 대응표본 t-검정을 통해 확인했다.

이에 주목해서 각 특약의 가입자들이 어떤 특징을 가지고 있는지 보다 구체적으로 파악하기 위해 연관분석을 진행했다. 연관분석은 장바구니 분석이라고도 불리며, 비지도학습법의 일종으로 자료에 있

는 항목(item)들간의 연관을 찾는 방법이다. 범주형 자료만으로 구성된 데이터이기 때문에 연관분석으로 유의미한 정보를 얻을 수 있고 특약 가입과 상호 연관성이 높은 고객의 특징을 파악할 수 있을 것이라 판단했다. 연관분석 결과를 나타내는 지표 중 향상도(lift)를 고려했다. 향상도($\frac{P(B|A)}{P(B)}$)는 A를 보유할 때 B를 보유할 확률이 해당 조건이 없이 B를 보유할 확률에 비해 얼마나 큰지에 관한 지표로, 값이 1보다 클수록 연관이 크다. 여기서 선행되는 조건인 A를 특약을 제외한 고객의 특징으로, 후행되는 결과인 B를 영상기록장치특약 및 마일리지 특약으로 적용해 특약 가입과 연관규칙을 갖는 고객의 특징을 파악하는 방향으로 분석을 진행했다.

5.2 연관분석 결과

Apriori 알고리즘을 통해 영상기록장치특약에 대한 연관분석 결과, 향상도(lift) 가 1.5 이상인 순서쌍을 추출했다. 향상도가 높은 순서쌍만 추출하여 특약 가입자들만 가지는 유효한 특징을 선별했다. 해당 변수 중 가입자만 가지는 특징을 선별하기 위해 미가입자와 공통적으로 가지는 특징을 제거하는 과정을 거쳤다.

피보험자연령대_60	피보험자연령대_50
차량경과년수_1	차량경과년수_2

(표4) 영상기록장치특약 가입자 특징

마찬가지로 마일리지할인특약에 대한 연관분석도 영상기록장치특약과 같은 과정을 거쳐 마일리지할인특약 가입자와 관련된 특징을 파악했다.

피보험자연령대_50	운전자한정특별약관코드_5
차량경과년수_3	차종_3

(표5) 마일리지할인특약 가입자 특징

결론적으로 영상기록장치특약은 50-60대, 차량경과년수 10년이하(신차 제외)인 고객에게 추천할 수 있다. 한편 마일리지할인특약의 경우 50대, 배기량 2000cc초과, 차량경과년수 5년 이상 10년 미만, 운전자한정특별약관이 '가족한정(형제자매제외)'인 고객에게 추천할 수 있다. 특약 가입자의 특징을 가진 미가입자가 특약에 가입한다고 가정했을 때 실제로 예상사고율이 감소한 데이터가 많이 관측되는 것을 확인했다.

	특약 종류	**가입 전 예측사고율**	**가입 후 예측사고율**
고객1	영상기록장치특약	3.495705	2.914301
고객2	영상기록장치특약	2.370356	1.550106
고객3	마일리지할인특약	2.177939	1.119478

(표6) 특약 종류에 따른 예측사고율 감소 예시

5.3 사고율 변화 검정

연관분석을 통해 파악한 특징을 가진 고객들이 특약 가입 전에 비해 가입 후에 사고율이 감소한다는 것을 확인할 수 있었다. 고객들의 사고율 감소가 통계적으로 유의한 것인지 검토하기 위해 대응표본 t-검정으로 확인했다. 모든 고객의 사고율 변화와, 연관분석을 이용해 추출한 고객의 사고율 변화가 통계적으로 유의한지 대응표본 t-검정을 통해 확인해보았다. 대응표본 t-검정(paired t-test)은 동일한 객체에 대해서 처치 전과 후에 측정된 자료 간에 평균의 차이가 있는지 검정하는 방법이다.

우선적으로 모든 영상기록장치특약 미가입자가 가입으로 전환했다고 했을 때 특약 가입 전보다 가입 후가 적은지 검정했다.

H_0: 특약 가입 전과 후 사고율에 차이가 없다. **H_1: 특약 가입 전보다 가입 후 사고율이 작다.**	
검정통계량	**p-value**
-0.35338	0.7238

(표7) 영상기록장치특약 가입 전후 대응표본 t-검정 결과

p-value > 0.1로, 유의수준 0.1 하에서 H_0를 기각할 수 없다. 즉, 영상기록장치특약 가입 전에 비해 가입 후 사고율이 유의하게 감소한다고 할 수 없다.

그러나 연관분석에서 파악한 특징을 가진 가입자만을 선별하여 이들이 특약 가입으로 전환했다고 했을 때, 특약 가입 전보다 가입 후 사고율이 유의미하게 감소한다는 결과가 나타났다.

H_0: 특약 가입 전과 후 사고율에 차이가 없다. **H_1: 특약 가입 전보다 가입 후 사고율이 작다.**	
검정통계량	**p-value**
2.1745	0.01485

(표8) 영상기록장치특약 가입자 특징 보유 미가입자 가입 전후 대응표본 t-검정 결과

p-value < 0.1로, 유의수준 0.1 하에서 H_0를 기각할 수 있다. 즉, 영상기록장치특약 가입 전에 비해 가입 후 사고율이 유의하게 감소한다고 할 수 있다.

모든 미가입자를 대상으로 사고율 변화를 관측했을 때는 특약 가입 전과 후에 차이가 없는 것처럼 비춰지지만, 가입자의 특성을 가진 미가입자를 선별해 가입 전과 후를 보면 유의미한 사고율 감소가 관측되는 것을 확인할 수 있었다. 따라서 가입자의 특징을 가진 미가입자들에 대한 타겟 마케팅을 통해 사고율 감소 효과를 기대할 수 있을 것이다.

마일리지할인특약 가입자의 특징을 가진 미가입자를 가입으로 전환한 경우, 사고율 차이도 다음과 같이 검정했다.

H_0: 특약 가입 전과 후 사고율에 차이가 없다. H_1: 특약 가입 전보다 가입 후 사고율이 작다.	
검정통계량	**p-value**
1.4107	0.07919

(표9) 마일리지할인특약 가입자 특징 보유 미가입자 가입 전후 대응표본 t-검정 결과

p-value < 0.1로, 유의수준 0.1 하에서 H_0를 기각한다. 즉, 마일리지할인특약 가입 전에 비해 가입 후 사고율이 유의하게 감소한다고 할 수 있다.

6. 시사점 제시

6.1 분석 방안 활용 아이디어

사고율 예측 모델링과 연관분석을 이용해 각 고객의 특징과 사고 위험률을 파악할 수 있다는 것을 확인했다. 이를 통해 보험사에서 각 고객에 대한 선제 대응이 가능할 것이라 판단했다. 분석 과정에서 얻은 결과를 이용해 고객의 사고 관리와 사고 예방 조치를 통한 손해율 관리에 도움이 될 보험사의 대응으로 특약 추천 서비스와 신규 고객 언더라이팅을 제안한다.

6.1.1 고객 맞춤 특약 추천 마케팅

1) 아이디어 내용

기존 고객 데이터를 활용한 고객 맞춤 특약 추천 마케팅을 제안한다. 맞춤형 특약 추천을 위해서 연관분석에서 파악한 기존 특약 가입자들의 특징을 고려해 특약을 가입할 의사가 높은 고객을 찾아내고 사고 예측 모델링의 요인 분석 결과를 이용해 사고율을 낮출 수 있는 고객에게 맞춤형 특약을 추천한다. 해당 분석 과정에서는 제공받은 데이터에서 영상기록장치특약과 마일리지 특약에 대해서만 분석할 수밖에 없었지만 다른 특약에 대해서도 적용해 초개인화된 추천 마케팅이 가능해질 것으로 생각한다. 더 나아가, 보험사에서 고객의 기본적인 정보뿐 아니라 고객의 직업, 라이프 스타일에 관해 파악할 수 있다면 운전자의 특성을 더욱 세분화하여 초개인화된 보험 특약 추천 마케팅이 가능해질 것이다.

2) 기대효과

위와 같은 과정을 통해 초개인화된 특약 추천 마케팅을 진행한다면, 가입 의사가 높을 것 같은 특정 고객을 타겟으로 한 마케팅이 가능하기 때문에 마케팅 비용 절감 효과를 기대할 수 있다. 또한, 영상기록장치특약과 같이 다수의 고객에게는 사고율 완화로 이어지는 가능성이 적은 특약의 경우, 전체 고객 중 특약 가입시 사고율 완화 효과가 클 것으로 예측되는 일부만을 선별해 추천함으로써 실질적인 사고율 감소 효과를 기대할 수 있을 것이다.

6.1.2 신규 고객 언더라이팅 고도화

1) 아이디어 내용

기존 고객의 데이터를 기반으로 학습시킨 사고율 예측 모델을 활용해 언더라이팅 과정을 고도화하는 것을 제안한다. 사고율 예측 모델링을 언더라이팅 과정에 활용한다면 기존 심사에 고려되었던 사고 이력 이외에도 다양한 고객 정보를 다각적인 측면에서 고려할 수 있을 것이다. 실제로 보험사 컨설팅 기업 Milliman은 고객의 처방전 기록 데이터를 이용해 사망률을 예측하고 사망률과 관련된 고객의 특징을 파악한 후 이를 기반으로 리스크 점수를 산출했다. Milliman은 산출한 리스크 점수가 임계점보다 미만인 경우에만 계약을 인수하는 방식으로 언더라이팅 과정을 고도화하여 보험사들의 순이익을 증가시키는 성과를 달성했다.

사고율 예측 모델링은 단순히 사고율을 예측하는 것에 그치지 않고 고객의 사고 발생률을 높이는 위험요인과 사고 발생률을 낮추는 완화요인을 파악할 수 있다. 위험요인과 완화요인처럼 사고 예측 모델링에서 파악한 사고 발생과 관련된 요인들을 기반으로 리스크 점수를 산출해 보험 가입 승인 기준에 반영한다면 계약 인수 심사 고도화가 가능해지고 정교한 손해율 관리가 가능해질 것이다.

또한, 사고율 예측 결과를 언더라이터가 쉽게 활용할 수 있도록 아래와 같은 '대시보드'를 제안한다. 대시보드는 사고율 예측 결과와 함께 보험 가입자의 성별, 연령대, 사고건수 등을 종합적으로 보며 빠르고 간편하게 심사를 진행할 수 있도록 구성했다. SHAP을 통해 도출한 고객의 사고 요인을 대시보드에 포함시켜 위험요인과 완화 요인을 제시하고 예측된 사고율에 대한 근거로 고려할 수 있을 것이다.

(그림11) 자동차보험 사고율 분석 대시보드 예상도

2) 기대효과

SHAP을 통해 보험 인수 심사 과정에서 심사 결과의 직관적인 근거를 고객에게 제시하는 것은 신뢰성을 제고하고, 계약 심사 결과를 고객에게 납득할 수 있도록 설명하는 데에 큰 도움이 될 것으로 기대된다. 대시보드를 통해 보험사에서는 언더라이팅 시간 단축 효과를 기대할 수 있다. 기존 보험 계약 심사 과정에서는 가벼운 사고일지라도 심사자가 사고 건수를 일일이 확인해 계약 승인을 결정해야 했기에 고객 입장에서는 심사 대기시간이 길어질 수밖에 없었다. 그러나, 머신러닝을 이용한 사고율 예측 결과가 반영된 대시보드와 리스크 점수를 기반으로 가입 승인 여부를 판별한다면 승인 과정이 단순화되고, 결과적으로는 심사 소요시간이 줄어들어 고객의 만족감을 높일 수 있을 것으로 기대된다.

6.2 제언

6.2.1 모델의 한계 및 발전 가능성

범주형 변수로만 구성된 데이터로 구축된 모델이기 때문에 수치형 변수들을 추가적으로 활용할 수 있다면 성능 개선이 가능할 것으로 예상된다. 또한, 개인 혹은 기업임을 파악할 수 있는 변수가 존재해 분리시킬 수 있다면 극단적으로 높은 사고율까지도 예측이 가능해져 보다 정교화된 사고율 예측이 가능할 것이다.

또한, 운전자 관련 변수들을 추가적으로 활용할 수 있다면 고객별 특성 및 사고율과 관련된 패턴을 더욱 세부화할 수 있을 것이다. 예를 들어 고객의 거주지와 직장지, 직업, 소득 구간, 면허 취득 연수 등과 같은 변수가 주어진다면 이를 바탕으로 고객의 특성 및 운전 숙련도 등을 보다 구체적으로 파악할 수 있을 것이다. 또한, 과거 사고 기록이 존재하는 운전자의 경우, 해당 사고 상황에 대한 구체적인 변수가 주어진다면 시공간적 정보를 반영한 모델링이 가능해질 것이다. 이처럼 보다 세분화된 고객 정보가 추가적으로 주어진다면, 더욱 정교화된 예측 모델링과 특약 추천 서비스의 품질 향상을 기대할 수 있을 것이다.

6.2.2 IoT를 활용한 사고율 예측 모델

위에서 제시한 모델의 한계를 극복하기 위해 분석 결과에 IoT를 접목시켜 추가 정보 수집을 고려할 수 있을 것이다. XAI를 기반으로 사고발생률 완화 요인을 선별한 결과, 피보험자연령대와 가입경력 코드가 가장 높은 중요도를 보였다. 연령대는 높을수록, 가입경력은 짧을수록 사고율이 높은 경향을 보였다. 이 결과를 바탕으로 IoT(Internet of Things) 관련 기술에 접목시켜 첨단 서비스나 고도화된 모델을 제공할 수 있을 것이라 생각한다. IoT를 통해 운전자의 비정상행동 감지 횟수, 구간별 운전 속도 등과 같이 추가적으로 수집된 운전 습관 데이터를 활용해 사고율 예측 성능을 향상할 수 있을 것으로 기대한다.

참고문헌

- 기승도. (2023). 사고감소를 위한 자동차보험제도. KIRI 리포트 포커스.
- 김수영 and 송종우. (2012). POT방법론을 이용한 자동차보험 손해율 추정. 응용통계연구, 25(1), 101-114.
- 최창희,홍민지. (2019). 빅데이터 활용 현황과 개선 방안. KIRI 보험연구원.
- BizWatch. (2023, 4월 29일). [보푸라기]'보험 문지기' 언더라이터의 세계. BizWatch. http://news.bizwatch.co.kr/article/finance/2023/04/29/0001
- 경향신문. (2023, 9월 11일). 자동차보험 상반기 손익 11% 감소…이동량 증가로 손해율 상승. 경향신문. https://m.khan.co.kr/economy/finance/article/202309111214001#c2b
- 뉴데일리경제. (2014, 3월 12일). 자동차보험 '블랙박스 특약' 할인률 낮아진다. 뉴데일리경제. https://biz.newdaily.co.kr/site/data/html/2014/03/12/2014031210026.html
- 보험매일. (2023, 5월 12일). 차보험 시장 할인 특약 경쟁 갈수록 심화. 보험매일. http://www.fins.co.kr/news/articleView.html?idxno=96712
- 최양호, 이우주, 오승철 and 이동환. (2014). 일반화선형모형을 이용한 생명보험 지급금 분석: 암 발생 사고자료를 중심으로. Journal of The Korean Data Analysis Society, 16(6), 3093-3106.
- 김서연, 이지윤, 목충협, 김상훈, 문석호, 김성범, ... & 경윤영. (2021). 해석 가능한 딥러닝 모델을 활용한 제조 공정의 불량 이상 감지. 한국경영과학회 학술대회논문집, 630-652.
- 정중영, & 강중철. (2006). 자동차보험 손해율에 관한 연구. Journal of The Korean Data Analysis Society, 8(6), 2445-2456.