삼성화재 자동차보험 고객 사고율 예측

💻 프로젝트 소개

한 줄 소개	로지스틱 회귀, 앙상블 트리 모델을 이용한 자동차보험 고객 사고율 예측 및 feature 발굴
분석 기간	2023.09 ~ 2023.11
관련 활동	http://www.riskds.com/
분석 도구	Python, R
수상	대상 (1등상)

1. 분석 목적

분석 배경 : 보험사에서 자동차보험 상품의 손해율 관리 및 매출 확보 강화 목적으로 **사고 위험이 적은 우량 고객**을 **확보**를 위한 다양한 **할인 특약 상품**을 개발하는 상황, **우량 고객**을 **선제적**으로 **확보**하기 위해 **우량 고객**의 **특성**이 반영한 **할인 특약 상품 개발**이 **필요**

분석 목적 : 자동차보험 가입 고객의 **사고율 예측 모델링**으로 **우량 고객 특징**을 **파악**해 **할인 특약 상품**에 관한 **인사이트**를 **도출**하기 위함

2. 분석 과정

▼ **활용 데이터**

삼성화재 자동차보험 가입 고객 데이터 (삼성화재 제공)

1. **데이터 전처리**

- 각 범주형 자료의 특성에 맞는 Labeling과 이상치 제거 작업을 통한 데이터 전처리 진행
- 데이터의 80%를 학습 데이터로, 20%는 검증 데이터로 사용하기 위해 데이터 분리

2. **EDA**

- 고객 특징 관련 EDA 및 상관관계 파악

3. **로지스틱 회귀 분석**

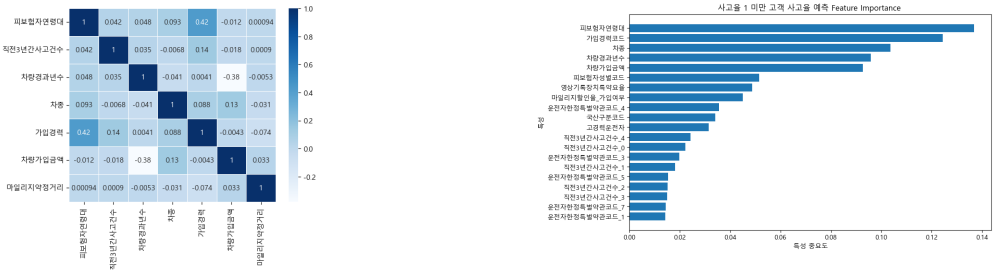
- 사고율의 분포가 0과 1 사이에 다수 분포되어 있어 높은 사고율을 예측하지 못 하는 문제를 해결하기 위해 진행
- 사고율이 1 이상인 고객을 고위험군 고객, 1 미만인 고객을 저위험군 고객으로 이진 분류 진행 및 우량 고객 특징 파악

4. **앙상블 트리 모델**

- Gradient Boosting, Decision Tree, Random Forest 모델을 결합한 앙상블 트리 모델을 이용해 고위험군 고객, 저위험군 고객의 사고율을 각각 예측
- Feature Importance, SHAP Value를 통해 우량 고객 특징 파악

3. 분석 결과

- 분석 결과



(왼쪽) 고객 특징 간 상관관계 / (오른쪽) 저위험군 고객 사고율 예측 모델링 Feature Importance (데이터 보안 문제로 구체적인 수치 공개 불가능)

- 로지스틱 회귀 분석 결과, 저위험군/고위험군 고객 분류 예측 성능은 정확도를 기준으로 0.62, Macro f1-score 기준으로 0.52으로 도출

- 각 고객군에 대해 앙상블 트리 모델을 이용해 사고율을 예측한 결과, 사고율 예측 성능은 rmse 기준 0.61로 도출

- 로지스틱 회귀 분석 결과와 앙상블 트리 모델의 결과를 종합해보면 직전3년간사고건수, 가입경력, 피보험자연령대가 사고율 예측에 기여도가 높은 고객 특성인 것을 확인, 피보험자연령대는 가입경력과 높은 양의 상관관계를 가져 가입경력과 함께 사고율 예측에 기여도가 높은 것으로 파악, 직전3년간사고건수가 적을 수록, 가입경력이 높을 수록 사고율이 낮아지는 상관관계를 보이는 것을 확인, 특약 상품 중 주행거리가 적은 마일리지 특약 상품을 가입한 고객일 수록, 운전자한정특별약관 중 부부 가입 혹은 기명피보험자1인한정을 가입한 고객일 수록 사고율이 낮은 것을 확인

- 결론

우량 고객 특징	**결론**	**제안**
- 직전3년간사고건수 **- 가입경력** **- 마일리지 특약 상품 가입**	사고율과 높은 상관관계를 보이는 것은 차량 요인, 운전자의 인구학적 요인이 아닌 **운전자의 운전 경향**과 관련된 요인들 → **운전**이 **능숙**하거나 **안전운전**을 하는 운전자/**주행량**이 **적어** 사고의 위험이 낮은 운전자가 **사고율이 낮은 고객들**	- 운전자의 **안전운전**을 **유도**하는 다양한 **BBI(주행습관기반) 특약 상품 개발** ex) 주행 속도, 급정거 횟수, 속도 준수 여부 등을 활용 - **자동차 주행**을 **줄이도록 유도**하는 **특약 상품 개발** ex) 대중교통 및 자전거 이용 빈도를 기반으로 보험료 할인 강화