질문

보험료 UPDATE는 언제?
모델 작동원리, feature importance,pdp,ice 원리
보험사의 마이데이터는 아직 덜 도입했다고함
https://news.bizwatch.co.kr/article/finance/2023/07/15/0001

##

***모든 부분에 어려웠던 점+해결방안 포함하기**
-> 보고서 [문제상황 및 해결 아이디어, p42] 보고 각자 맞는 파트에 같이 설명하기
EX.
목차 [분석과정]-> 데이터 수의 한계로 인해, 각 보험별 보장내용을 기반으로 총 15개 모델을 못만듬(Underfitting의 가능성)-> 대신, 3가지로 예측모형을 만들고 각 보장내용을 변수로 사용함

분석 배경 및 목적

[기존 상황]

1. 초기 보험료 산정 시-> 고객이 기입한 신상정보에 의존,
   이후 사고율 산정-> "사고건수/계약건수", '보험금'/'보험료' 등 정적인 방식

=> 사고율에 영향을 줄 수 있는 고객의 새로운 패턴,동향 파악이 X(개인의 사고율 예측을 위해서는 생활, 소비, 패턴 등 수 많은 요인이 존재하지만 기존 방식은 나이,성별,보험가입경력 등 제한적인 신상 데이터만을 이용해 그 정확도가 낮다. +설명력의 부족

[마이데이터의 등장]

제공해준 카드 데이터-> 개별 고객의 생활 전반 정보를 폭넓게 이해할 수 있어 개인화된 예측이 가능할 것이라 예상

[분석 목적]

#보험사의 손실 방지(회사 전체 수익 극대화)
수지상등의 원칙 지키기->보험료 산출과정에서 정확한 사고율 예측

1.보험료 설정 시 해당 요인을 집중적으로 파악(예측 모형 만들 시 반영, 모형의 무게 간소화), 초기 설문조사 시 고객의 신상조사에서 해당 요인을 더 정확, 집중적으로 파악

#고객관리(고객 만족도 상승)
사고율에 영향을 끼치는 요인 분석(설명가능성)
1.복잡한 공식에 의해 산출된 보험료가 사고율을 바탕으로 어떤 근거로 책정되었는지 설명-> 보험상품과 보험료에 대한 신뢰(신규 고객 유입+고객이탈 방지)
2. 보험료 인상/인하, 프로모션, 집중적 관리 (고객 이탈 방지)

분석 목적을 바탕으로 **[가설 설정]**
(1) 사고율 예측과정에서 보험개발원 데이터+카드데이터(마이데이터)가 추가되면, 더 정확하 사고율 예측이 가능할 것이다 => 보험별 "마이데이터의 실효성"에 대한 판단

(2) 정확히 예측된 사고율 모형에 중요한 영향을 끼치는 변수와 그 영향력을 파악할 수 있다면 더 정확한 보험료를 산출

[검증 전략= 분석 과정]

2가지 중시 사항 앞에다 박고 시작(P18)
1) 모든 분석 및 모델링 과정~
2) 단순 AI~
*좋은 모형이란? 예측력+설명가능성-> 특정 고객의 사고율이 높게 측정되었다면 어떤 요인으로 인해 해당결과가 산출되었는지 설명할 수 있어야함(참고)

(가설1) 보험명세서 데이터만을 이용한 사고율 모형보다 보험명세서+카드데이터를 함께 이용한 사고율 모형의 성능이 더 좋음을 확인-> 해석의 용이함을 위해 3가지 보험(생명,장기손해,자동차) 종류에 따라 각기 다른 사고율 예측 모형
(1-1) IDEA_제시된 사고율로 종속변수로 설정하면 (NULL,결측치,0) 값으로 모델 학습이 제대로 이루어지지 않음 -> 보험종류별 다른 전략 구사
#생명보험
대부분 0값이지만 결측치는 단 10개-> 'ACCD_CUNT' 변수를 활용하여 변수 해석
#장기손해보험 및 자동차보험
보고서에 나와있는대로(확률값으로 사고율 비교)
예상 사고율(보험데이터만 사용) VS 실제 사고율(카드데이터 함께 사용)
+ 로지스틱 회귀의 기술적인 측면 설명

(가설2) 3가지 보험에 대한 예측 모델링 후, 요인 파악+군집화(K-MEANS)
보험별 중요변수 해석 후 risk_code 별로도 해석함(boxplot을 통해서)

원칙: 사람이 임의적 개입이 x, 머신러닝 알고리즘 모델이 스스로 파악!-> feature importnace

#생명보험
이 경우, 사실 가설1은 안하고, 가설2만 함,
Undersampling 후 최대한 설명력과 예측력이 높은 모델로 요인 해석, 로지스틱 회귀 사용,

#장기 손해 보험 및 자동차 보험
해석 시 중요변수의 기준은?
feature importance를 scree plot으로 시각화 후, 중요도가 갑자기 감소하는 부분!
ice plot(+pdp)으로 변수 영향 해석

***기술->보고서 해당 부분 참고_간소하게..(어차피 모름)**

[전처리]

간단하게, 파생변수 뭐썼는지, 범주형 변수의 원핫 인코딩, 변수 삭제 등등

[예측 모델링 과정]

각 보험별로 구체적으로 설명x-> 전체적인 prcess(이런 과정을 거쳤다는 거 보여주기
생명보험
undersampling 처리, RFE로 변수 선택(데이터 수 적기 땜에, 오버피팅 막으려고), 로지스틱 회귀 왜 썼는지

후보 모델로 이러이러한게 있었고, 왜 그 모델을 선정했는지
k-fold를 썻는데, 이게 뭔지, 왜 썻는지
학습평가
평가지표는 무엇을 썻는지..

[해석 과정+ 군집화]

feature importance, ice(pdp) 예시 들면서 간략하게 소개(좀 더 집중)

[분석 결과]

(보고서 p22~p37)

<가설에 대한 검정 결과>
생명보험은 가설2만 검정

나머지는 가설1,2검정-> 장기손해보험은 가설채택

군집화 분석 결과는 [적용방안 및 기대효과]에 넣는 게 좋을듯

[적용 방안 및 기대효과(금융권 내 효익)]->좀 더 구체화하기

보고서(p37~41)

페르소나 설정

+김필상 웹사이트제작(큐알코드)

[느낀점]

보고서(p45~46)

[차별성, 강조하고 싶은 것]

내용 중복되도 좋으니까 강조하고 싶은거 한번 더 강조