Model Development

Project - 1

보험 사기 탐지 모델 개발

배경

- AI 시대의 변화 속도가 매우 빠르기 때문에, 이에 발맞추는 것이 중요하다고 생각합니다.
- 보험업계에서는 항상 보험 사기가 주요 이슈이며, 지속적으로 손실을 입고 있습니다.
- 데이터 분석을 활용하여 보험 사기 리스크를 효과적으로 줄이고자 하였습니다.

*해당 분석은 실제 데이터가 아닌 **연습용 데이터**로 제작되었습니다.

작업 기간

2025. 02. 24 - 03. 31

구성원

기여도 100%

데이터분석 기술 스택

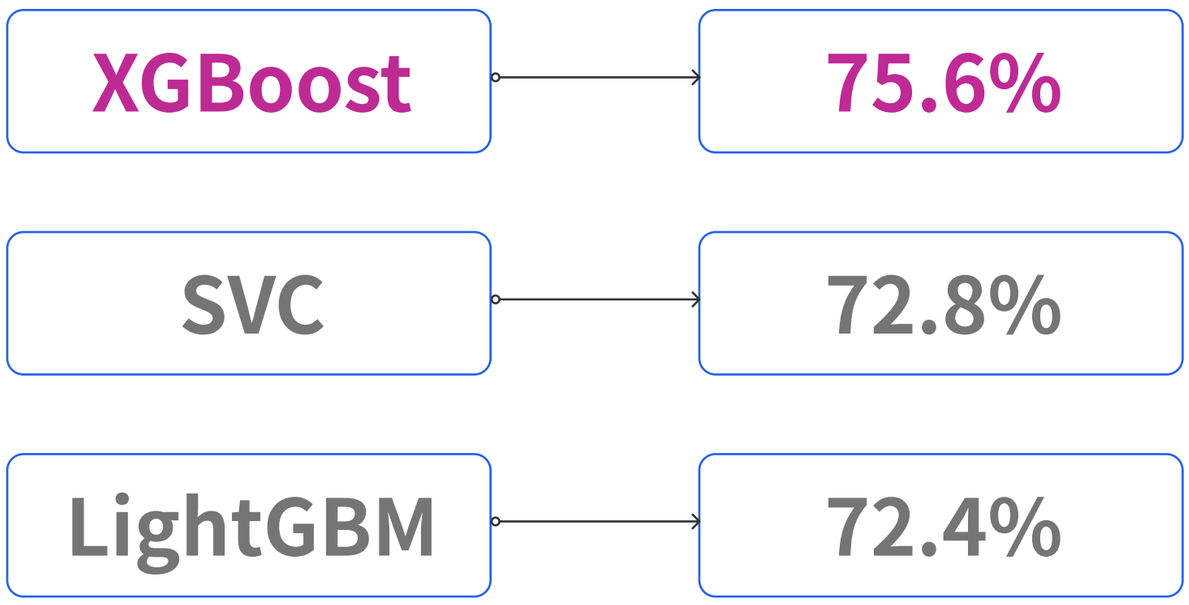
프로젝트 요약

- 사기 리스크 예측 모델(XGBoost) 정확도 75.6%
- 주요 변수: 사고 유형, 피해금액
- 산점도·Feature Importance로 시각화
- 보험 사기 조기 탐지 기준 제시

Model Development

Project - 1

모델별 정확도 결과



XGBoost Classifier 분석

- **모델 선택 이유**
  - 부스팅 기법을 활용한 분류 모델이며, 대규모 데이터셋에서 높은 예측 및 과적합 방지하기에 해당 모델 선택
  - 보험 사기 탐지 문제에서 연속형 및 범주형 번수 혼합된 데이터셋을 효과적으로 처리할 수 있음

- **세부사항**
  - XGBClassifier() 기본 설정을 사용하여 훈련 진행.
  - 해당 모델을 활용해 75.6%의 정확도를 얻었으며, SVC 모델보다 정확도가 높음
  - 부스팅 기법을 사용해 데이터의 패턴을 정밀하게 학습하여 Decision Tree 기반 모델보다 높은 성능을 보임
  - 데이터의 불균형이나 이상치가 있는 경우 모델의 성능이 영향을 받을 수 있음
  - ✅ **적용 모델 중 가장 정확도 높은 모델**

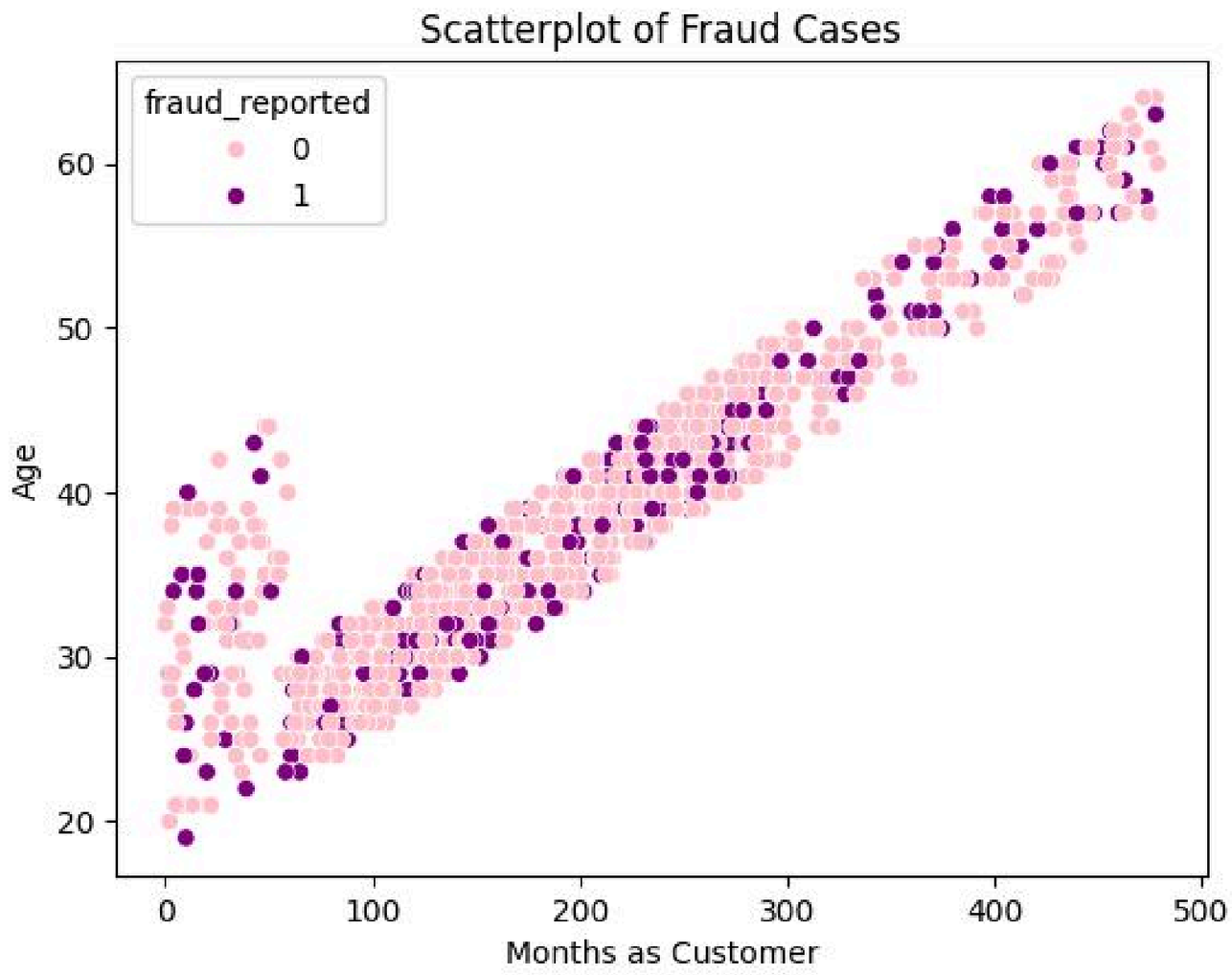
LightGBM 분석

- **모델 선택 이유**
  - XGBoost와 유사한 부스팅 기법을 사용하지만, 대량의 데이터를 빠르게 학습할 수 있음
  - 빠른 학습 속도, 과적합 방지 기능, 대규모 데이터셋 처리에 최적화

- **세부사항**
  - 트리의 깊이를 2로 제한하여 과적합 방지 및 10개의 부스팅 스텝을 사용하여 학습 진행
  - 해당 모델 72.4%의 정확도를 얻었지만, XGBoost의 정확도 보다 낮게 나옴
  - 이후 하이퍼파라미터를 최적화하면 성능 향상이 가능함을 확인
  - 데이터 전처리 개선을 통해 (이상치 제거, 범주형 변수 인코딩 방식 변경) 성능 향상 가능

Model Development

Project - 1



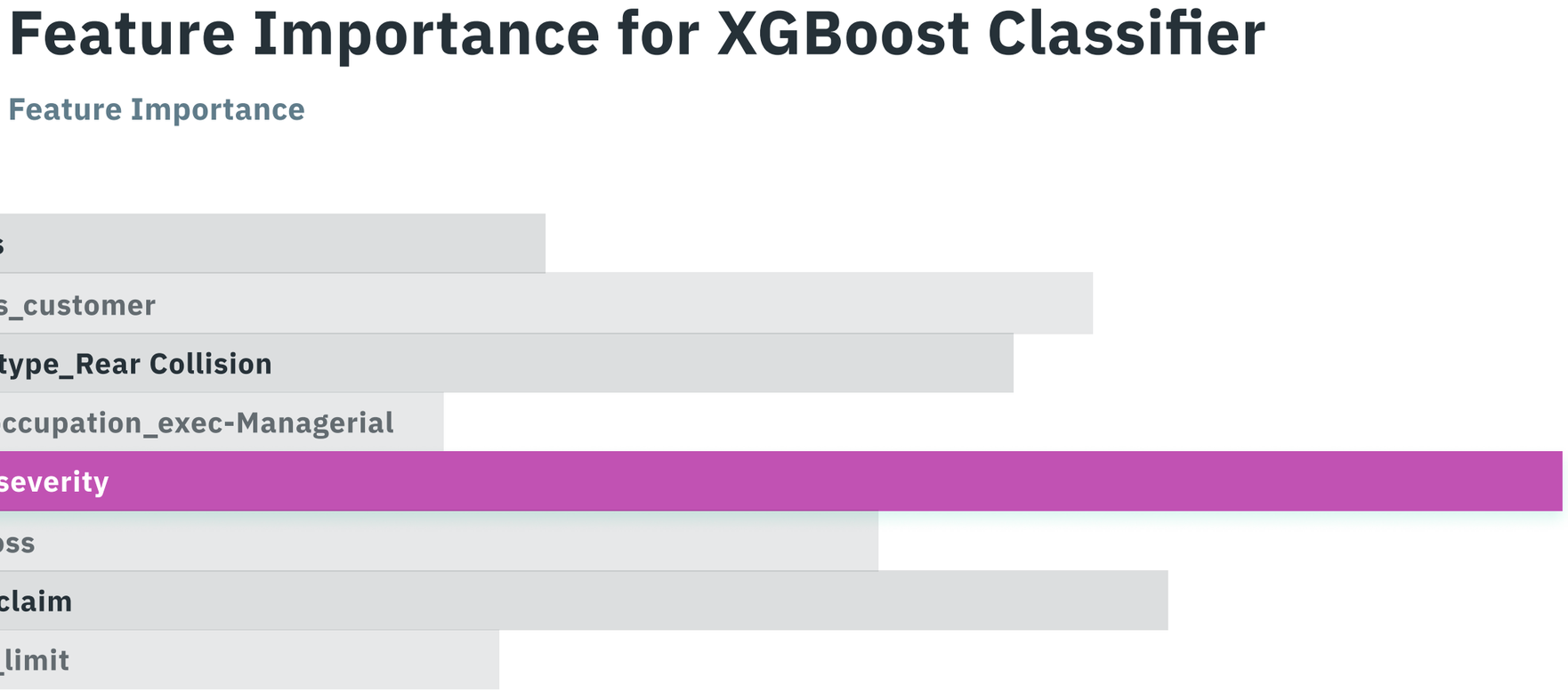
사기 X

사기 O

사기 의심 사건 여부 산점도

- 20-40대 사이 고객들은 주로 보험 사기 신고가 많이 발생하는 경향이 있음
- 가입 기간이 길어질수록 사기 신고 비율이 감소함
- 나이가 많아지고, 가입 기간이 길어지면 보험 사기 신고 비율이 현저히 낮음

- ✅ **보험 사기는 젊고 가입 기간이 짧을 때 발생 확률이 높다.**



- 결과적으로, 해당 변수들이 보험 사기 예측에 영향을 받았는 지 알 수 있었습니다.
- ✅ **사고의 심각도와 피해금액**이 보험 사기 예측에서 가장 중요한 변수로 작용하는 것을 도출하였습니다.
- 이후, 사기 탐지 기준을 구체화하여 **고위험군 고객을 조기 발견하는 시스템**을 구축하고자 합니다.