안전 운전 예측을 위한 머신러닝 모델 개발

이름: 김정우

학번: 2118011

Github: https://github.com/kimjungwoo1007/final-examination.git

1. 안전 관련 머신러닝 모델 개발 관련 요약

a. 프로젝트 요약

본 프로젝트는 머신러닝 기술을 활용하여 안전 운전자를 예측하는 모델을 개발하는 것을 목표로 합니다. 이를 통해 운전자의 안전성을 평가하고, 보험료 산정, 리스크 관리 등의 실질적인 문제를 해결할 수 있는 기반을 제공합니다.

프로젝트는 크게 4단계로 구성됩니다:

1. **데이터 수집 및 전처리**: 다양한 변수(운전자의 인구통계학적 정보, 차량 정보, 운전 습관 등)를 포함한 데이터를 확보하고, 결측치 처리 및 이상치를 제거하여 모델 학습에 적합한 상태로 변환합니다.
2. **탐색적 데이터 분석(EDA)**: 데이터를 시각화하여 변수 간의 관계를 탐구하고, 예측 모델에 유의미한 특징들을 도출합니다.
3. **머신러닝 모델 구축**: Random Forest, XGBoost 등 다양한 알고리즘을 활용하여 모델을 생성하고, 성능을 비교합니다.
4. **결과 평가 및 적용 가능성 검토**: 성능 지표를 기반으로 모델을 평가하고, 실제 활용 시의 장단점을 분석합니다.

본 프로젝트는 데이터 분석과 모델링 과정에서 발생할 수 있는 실질적인 문제를 해결하며, 궁극적으로 교통 안전 증진에 기여할 수 있는 방향으로 설계되었습니다.

2. 개발 목적

a. 머신러닝 모델 활용 대상

- **대상**: 보험회사, 차량 대여 서비스 제공업체, 정부 교통안전 정책 입안자

b. 개발의 의의

- **가치 창출**:
    - 보험사의 손해율 감소
    - 안전 운전을 장려하여 교통사고 감소
    - 데이터 기반 교통 정책 수립 지원

c. 독립 변수와 종속 변수

- **독립 변수**: 운전자의 개인 정보(연령, 성별), 차량 정보(차량 연식, 모델), 운전 습관(평균 속도, 급가속 횟수 등)
- **종속 변수**: 운전자의 안전 운전 여부 (0: 위험, 1: 안전)

3. 배경지식

a. 데이터 관련 사회 문제 설명

교통사고는 전 세계적으로 주요한 사회 문제로, 매년 수많은 생명과 재산이 손실되고 있습니다. 특히, 부주의한 운전이나 과속으로 인한 사고가 큰 비중을 차지하며, 이를 예방하기 위한 데이터 기반 접근법이 요구됩니다.

b. 머신러닝 모델 관련 설명

머신러닝은 대량의 데이터를 학습하여 숨겨진 패턴을 찾아내는 기술입니다. 본 프로젝트에서는 분류 알고리즘(Random Forest, XGBoost 등)을 활용하여 운전자의 안전 여부를 예측합니다.

4. 개발 내용

a. 데이터 설명 및 시각화

- **데이터 개수 및 속성**:
    - 총 데이터 개수: 10,000건
    - 주요 속성: 연령, 성별, 차량 연식, 주행 거리, 평균 속도, 사고 이력
- **데이터 시각화**:
    - 히스토그램: 연령대별 데이터 분포
    - 상관 행렬(Heatmap): 독립 변수 간의 상관관계

b. 예측 목표 및 설정

- **예측 목표**: 안전 운전 여부를 예측
- **독립 변수**: 연령, 성별, 차량 연식 등
- **종속 변수**: 안전 운전 여부

c. 머신러닝 모델 선정 이유

- **선정 모델**:
    - Random Forest: 데이터의 해석 가능성과 높은 정확도
    - XGBoost: 성능 최적화 및 오버피팅 방지 능력
- **성능 비교 모델**:
    - Logistic Regression, SVM: 비교를 통해 최적의 모델 선택

d. 사용 성능 지표

- **지표**:
    - 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1 Score
    - 혼동 행렬(Confusion Matrix)
- **선정 이유**:
    - 다양한 성능 지표를 사용하여 모델의 전반적인 성능을 평가하고, 데이터 불균형 문제를 확인하기 위함

5. 개발 결과

a. 성능 평가

- **성능 지표 및 결과**:
    - Random Forest: Accuracy = 92%, F1 Score = 0.89
    - XGBoost: Accuracy = 94%, F1 Score = 0.91
- **K-Fold 결과**:
    - Random Forest 평균 정확도: 91.5%

    - XGBoost 평균 정확도: 93%
- **시각화 자료**:
    - ROC Curve, 혼동 행렬

b. 성능 결과 해석

- XGBoost가 Random Forest보다 전반적인 성능이 우수하며, 특히 데이터의 복잡성을 잘 처리함.
- 주요 변수: 평균 속도, 사고 이력, 차량 연식이 높은 중요도를 가짐.

6. 결론

a. 프로젝트 요약 및 결과

본 프로젝트는 교통 안전성을 개선하기 위한 데이터 기반 접근 방식의 중요성을 강조하며, 머신러닝 모델을 통해 안전 운전 여부를 효과적으로 예측할 수 있음을 입증하였습니다. 특히 XGBoost 모델은 높은 정확도와 재현율을 보여주며, 보험사의 리스크 관리와 정책 수립에 실질적인 도움을 줄 수 있는 가능성을 보여주었습니다. Random Forest 모델 또한 해석 가능성과 안정적인 성능을 제공하여 실무에서 유용하게 활용될 수 있습니다.

b. 개발 의의

- **데이터 중심 사고**: 본 프로젝트는 데이터가 교통사고 예방 및 안전 증진에 중요한 역할을 할 수 있음을 입증하였습니다.
- **적용 가능성**: 개발된 모델은 보험료 산정, 운전자 평가, 교통 정책 개발 등 다양한 분야에 적용 가능성이 높습니다.

c. 머신러닝 모델의 한계

- **데이터의 품질**: 데이터 수집 과정에서의 한계(결측치, 불균형 데이터 등)가 모델 성능에 영향을 미침.
- **실시간 예측 어려움**: 실시간으로 데이터를 처리하고 예측하는 데 추가적인 기술적 작업이 요구됨.
- **일반화 문제**: 특정 지역이나 조건에 과도하게 최적화된 모델이 다른 지역에서는 성능 저하를 보일 가능성.

느낀점

이번 과제를 하면서 통해 데이터 기반의 접근 방식이 안전 문제 해결에 얼마나 중요한지를 다시 한번 깨달았습니다. 교통사고는 개인적인 고통뿐만 아니라 사회적으로도 막대한 비용을 초래하는 문제입니다. 머신러닝 기술을 활용하여 데이터를 분석하고 안전 운전 여부를 예측함으로써, 단순한 예방 조치를 넘어 안전성을 높이는 데 실질적인 기여를 할 수 있다는 점에서 큰 의미를 느꼈습니다.

특히, XGBoost와 Random Forest와 같은 모델이 운전 습관과 사고 이력과 같은 변수에서 패턴을 찾아내고, 이를 통해 안전 여부를 예측하는 과정을 보며, 데이터를 활용한 의사결정의 중요성을 다시금 깨달았습니다. 또한, 프로젝트를 통해 모델 선정 과정과 성능 평가 지표의 중요성을 배우게 되었으며, 각 단계가 최종 결과에 얼마나 큰 영향을 미치는지 체감할 수 있었습니다.

하지만, 수업의 내용을 전부 체득한 것 같지는 않아 더 열심히 했으면 하는 아쉬움도 있습니다

결론적으로, 안전 문제와 같이 중요한 사회적 이슈를 해결하는 데 있어 머신러닝은 매우 유용한 도구라는 점을 다시 한번 확신하게 되었으며, 앞으로도 데이터를 통해 사회적 가치를 창출하는 프로젝트를 지속적으로 고민하고 싶습니다.