고령운전자 교통사고 심각도 분류 예측 모델 연구

line 1: Name
line 2: *dept. name of organization*
line 3: *name of organization*
line 4: *Student ID*

***Abstract*— 최근 고령 운전자 수가 증가하고 있으며 이에 대한 사회적 문제가 대두되고 있다. 본 연구에서는 2018 년부터 2023 년까지 서울시 데이터를 중심으로 노인 교통사고 심각도의 영향 요인에 대해 Boosting 과 Bagging 알고리즘을 사용한 몇 가지 모델을 통해 분류 분석하고 이에 따른 정책을 제안하고자 한다.**

***Keywords*— 고령운전자, 교통사고, 분류, CatBoost**

I. INTRODUCTION

1. 배경

전세계적으로 고령화가 가속화되면서 고령 운전자가 늘어나고, 이에 따른 고령 운전자 사고 또한 심각해지는 추세다. 이때 고령운전자는 도로교통공단에 따르면 2022 년 교통사고건수가 2021 년 대비 3.1% 감소한 반면, 고령운전자 교통사고 건수는 2021 년 대비 8.8% 증가하였다. [1] 또한 한국교통안전공단이 발표한 자료에 따르면, 전체 교통사고 중 고령 운전자가 낸 사고 비중이 2019 년 14.5%부터 지난 5 년간 꾸준히 증가하여 20%를 기록하였다. [2] 최근 국내에서도 고령 운전자의 교통사고가 잇달아 일어났으며, 고령 운전자로 인해 많은 사람들이 다치거나 사망한 사건이 발생하였다. 특히 올해 7 월 시청역에서 68 세 남성이 몰던 차량이 운전자 과실로 보행자를 덮쳐 9 명의 목숨을 앗아간 사건으로 인해 고령 운전자의 교통사고를 예방하기 위한 대책이 촉구되고 있다. [3]

다만, 고령운전자의 사고는 늘고 있으나 이에 대한 대책은 미흡한 상황이다. 현재 정부는 2018 년부터 고령 운전자의 사고 예방을 위해 운전면허를 자진 반납하도록 하는 정책을 시행하고 있으나, 경찰청에 따르면 2022 년 2.6%, 2023 년 2.4%로 매우 낮은 상태이다. [4] 따라서 이러한 국내 고령 운전자 교통사고 증가와 관련 사회적 논의가 활발해짐에 따라 고령운전자의 교통사고 심각도 분석을 주제로 선정하고자 한다.

2. 선행연구 검토 및 차별점

- 선행연구 검토

먼저, 고령운전자의 교통사고 심각도 분석을 위하여 선행연구를 검토하였다. 교통사고 심각도 분석에 대한 연구는 주로 전체 운전자 집단을 대상으로 이루어졌다. [5, 6] 교통사고 심각도 분석은 주로 회귀 분석과 분류 분석, 그리고 클러스터링으로 나뉜다. 먼저 교통사고 데이터셋을 바탕으로 KMeans 와 같은 클러스터링 기법을 활용하여 운전자 그룹 및 특성을 분석한 연구가 있었다. 클러스터링 결과를 바탕으로 각 운전자 그룹과 특성별 위험도를 산출하고 분석하였다. [7] 또한 한 연구는 2015 년부터 2019 년까지 총 1571 개의 교통사고 데이터와 기상 데이터를 활용하여 교통사고 사상자수에 대한 회귀 분석을 진행하였다. [8] 2015 년부터 2017 년까지 전체 운전자의 고속도로 교통사고 데이터를 대상으로 LightGBM 등의 5 가지 모델을 활용하여 교통사고 심각도를 분류하는 연구도 있다. 한국도로공사에서 사용하는 사고 피해 정도 등급에, D 등급을 추가하여 4 중 분류를 수행하였고, LightGBM, CatBoost 순의 성능을 보였다. 해당 연구를 통해 사고차량 수, 사고 유형, 사고지점, 사고차로 유형, 사고차량 유형이 교통사고 심각도 추정에 중요하다는 점을 확인하였다. [9, 10]

- 차별점

이처럼 선행연구를 살펴본 결과, 대부분의 연구가 주로 전체 운전자 그룹에 초점을 맞춰 분석을 수행하였다. 본 프로젝트에서는 기존의 연구와는 다르게 최근의 사회적 논의를 고려하여 고령 운전자에 초점을 맞춰 분석하는 것이 시의성 있다고 판단하여 해당 주제를 선정하였다. 또한 대부분의 연구가 비교적 과거의 데이터를 사용하고 있어 고령운전자가 늘어난 최근의 추세를 잘 반영하지 못하고 있다. 이에 따라 본 프로젝트에서 2018 년 ~ 2023 년의 실제 서울시 데이터를 활용하여 분석하는 것이 의미가 있을 것이라고 생각된다. 노인 운전자를 대상으로 심각도 분석 모델을 구축하고, 이 과정에서 교통사고 심각도에 영향을 미치는 요인들을 분석하여 고령운전자의 교통사고 증가 문제 해결을 위한 정책을 제안하고자 한다.

II. METHOD

- 데이터 검토

노인 교통사고 데이터는 도로교통공단 교통사고 분석 시스템(TAAS)에서 제공하는 2018 년부터 2023 년 사이 교통사고 자료를 기반으로 가공한 것이다. [11] 노인의 기준은 65 세 이상으로 하였고 해당 자료에서 가해운전자가 65 세 이상인 서울시 데이터만을 사용했으며 총 35,319 건의 사고데이터를 분석에 사용했다. 또한 직관적으로 차량의 통행 속도가 사고 심각도에 유의미한 연관 관계가 있을 거라 생각했지만 국내 선행 연구는 부족한 상태이다. 이는 도로교통공단 교통사고 분석 시스템(TAAS)에서 제공하는 데이터가 사고유형 - 과속이라든지, 제한속도 같은 데이터만 있을 뿐 개별 사고에서 사고직전 속도가 얼마였는지와 같은 데이터는 없기 때문이라고 추정된다. [12] 이를 대신해서 서울시 열린데이터 광장에서 가져온 월별 구별 평균속도를 추가했다.

- 데이터 전처리

모델링을 하기에 앞서 데이터 전처리를 수행하였다. 먼저 종속변수 설정을 위해 사고내용 컬럼 내용의 사망, 중상 사고를 1, 부상신고와 경상사고를 0 으로 매핑하여 사고 심각도 컬럼을 생성하였다. [13] 다음 독립변수 전처리를 위해 중복된 행과 결측 값을 확인하였다. 중복 행은 존재하지 않았고, 결측 값은 피해운전자 차종, 피해운전자 성별, 피해운전자 연령대, 피해운전자 상해정도 컬럼에 각 991 개씩 존재하였다. 다만, 결측 값을 확인해본 결과 이는 피해운전자가 존재하지 않는 사고였기에 해당 결측 값을 정보 없음 (차량 단독) 값으로 대체하였다.

시군구 컬럼은 시, 구, 동 컬럼으로 분할했으며 미분류 항목 하나를 확인하고 제외했다.

일과 공휴일 컬럼의 경우 사고번호 형식이 '20180101**********'와 같은 형식임을 확인하고 여기서 발생 년도 월 일자 데이터를 추출해 이를 기반으로 공휴일과 일 컬럼을 추가했다.

요일 컬럼을 그대로 쓰는 것보다 주말 여부로 파생변수를 추출하여 사용하는 것이 타겟 변수와 더 관련이 있을 것이라고 판단하였다. 따라서 요일 컬럼을 바탕으로 주말과 주중으로 나누는 주말 여부 컬럼을 추가하였다.

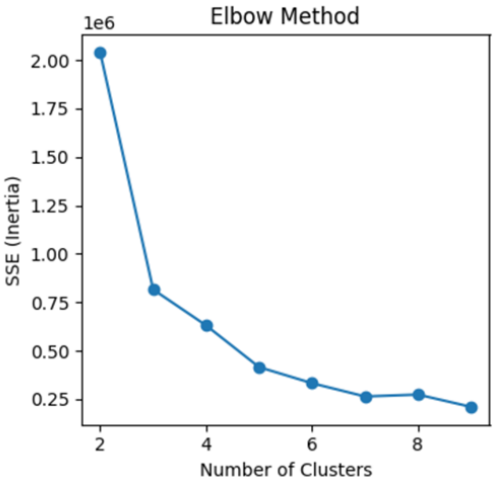
사고유형 컬럼의 경우 '-'을 기준으로 왼쪽에는 차대차, 차대사람, 차량단독이라는 포괄적인 정보와 오른쪽에는 충돌, 횡단중, 차도통행중 등 더 세부적인 정보가 포함되어 있었다. 사고유형은 매우 세분화되어 있기에, 이를 줄이기 위하여 '-'를 기준으로 사고유형 1 과 사고유형 2 로 구분하였다. 도로형태 컬럼 또한 '-'를 기준으로 포괄적인 정보 (단일로, 교차로, 기타, 주차장, 미분류)와 세분화된 정보 (교차로 안, 교차로 부근, 교량 등)를 포함하며 많은 범주를 가지고 있었다. 따라서 이를 '-'을 기준으로 도로형태 1 과 도로형태 2 라는 2 개의 컬럼으로 분리하였다. 도로형태 1 컬럼에서 미분류에 속하는 컬럼은 비교적 소수이기에 이를 기타 범주와 묶어 '기타/미분류'로 범주화하였다.

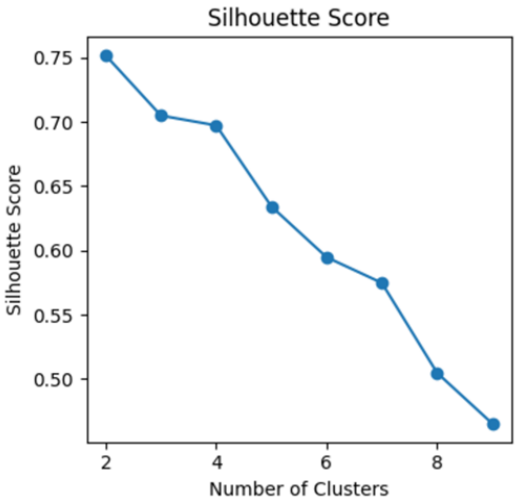
기상상태 컬럼은 맑음, 비, 흐림, 기타, 눈, 안개라는 총 6 가지의 범주로 이루어져 있었는데, 안개는 4 건으로 매우 적었기에 이를 흐림과 묶어 흐림, 안개로 재범주화하였다. 노면상태도 마찬가지로 7 가지의 범주 중 전체에서 1% 미만의 빈도를 가진 범주들 (적설, 해빙, 침수)이 있어 성격이 비슷한 다른 범주와 묶어 하나로 변경하였다. 적설, 해빙은 서리/결빙과 함께, 침수는 젖음/습기와 묶어 재범주화하였다.

피해운전자 연령대는 8 개의 연령 구간과 정보 없음, 기타 불명까지 총 10 개의 범주를 가지고 있었다. 81 건의 기타 불명은 피해운전자가 있으나 알려지지 않은 것이기에 최빈값인 31-40 세로 대체하였다.

'사고 번호'에서 연, 월, 일 특성을 추출해 내었다. '월', '일'은 시계열 데이터로 이들이 가진 주기성을 반영하기 위하여 푸리에 특징을 활용한 전처리를 수행하였다. 데이터를 사인 함수와 코사인 함수로 변환해 머신 러닝 모델이 패턴 학습을 보다 원활하도록 만들었다.

'동' 특성은 그 가지수가 457 개로 비슷한 특성을 지닌 동끼리 군집화하여 사고 심각도의 공간적 요인을 반영하고자 하였다. Cramér's V 계수를 통해 '동' 특성과 가장 상관성이 높은 변수를 확인해 본 결과 '사고유형 1', '도로형태 2'으로 나타났다. 최적의 군집 수 설정을 위해 Elbow Method, Silhouette Score 를 확인하여 최적의 군집 수를 3 으로 설정하였다. K-means 를 이용하여 군집화 진행 후 동 별로 각 군집 번호를 할당한 '동 군집' 변수를 생성하였다.





'가해운전자 차종' 과 '피해운전자 차종'에서 데이터가 100 개 미만인 값들은 '기타' 값으로 처리하였다. 그 결과 '가해운전자 차종'에서 '사륜오토바이(ATV)', '특수', '기타불명', '피해운전자 차종'에서 '농기계', '사륜오토바이(ATV)', '미분류', '특수', '기타불명'는 모두 '기타' 값으로 대치하였다.

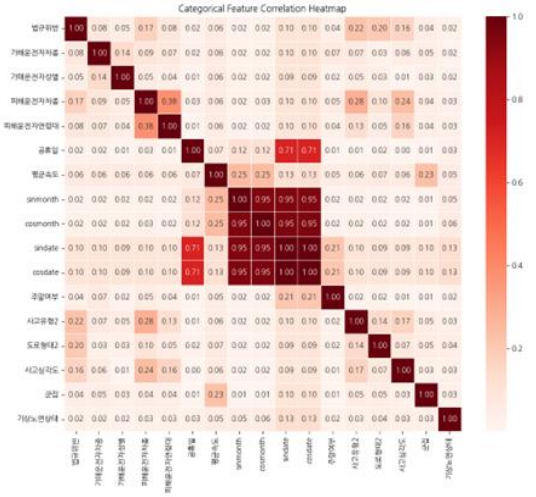
- 변수 선정

지금까지 생성된 feature 컬럼들 중에서 다른 feature 로 사용할 컬럼과 중복되는 컬럼, (사고번호, 발생년월, 발생년월일, 요일, 시군구, 시, 연, 동, 도로형태, 사고유형) 의미가 없어 사용하지 않을 컬럼들 (가해운전자 연령대, 피해운전자 성별)을 제거하였다. 더불어, 타겟 컬럼인 사고 심각도와 직접적으로 연결되어 성능에 영향을 주는 변수들 (사고내용, 사망자수, 중상자수, 경상자수, 부상신고자수,

가해운전자 상해정도, 피해운전자 상해정도) 또한 삭제하였다.

- 영향 변수 도출

전체 데이터 컬럼을 활용한 것에서 성능을 향상하기 위해 상관관계 분석을 활용하였다. 사용 데이터는 다양한 범주형 변수를 포함하고 있기에, 범주형 변수 간의 상관관계를 확인할 수 있는 크래머 V 상관계수를 활용하였다. 이때 크래머 V 상관계수는 카이 제곱 독립성 검정을 활용하여 계산하며, 0~1 의 값을 갖는다. 0 에 가까울수록 상관관계가 낮고, 1 에 가까울수록 두 변수 간의 강한 상관관계를 시사한다. 먼저, 분석을 통해 feature 간 상관관계가 높아 예측에 방해가 되는 다중 공선성 문제를 방지하고자 하였다. 두 컬럼 간 0.5 이상의 상관관계 값을 갖는 경우를 확인하였다. 기상상태와 노면상태는 0.58 의 상관계수를 나타내 이 둘을 '기상노면상태'라는 하나의 컬럼으로 결합하였다. 앞선 처리에서 2 개로 분리한 도로형태 1, 도로형태 2 와 사고유형 1, 사고유형 2 또한 높은 상관계수를 보였다. 이에 따라 비교적 포괄적인 정보를 가지고 있는 도로형태 1 과 사고유형 1 컬럼을 제거하였다. 또한 사고유형 1 은 피해운전자의 일부 컬럼과 높은 상관계수를 보였는데 이는 앞선 처리에서 사고유형 1 을 제거하면서 문제를 해결하였다. 더불어, 사고 심각도와 feature 간의 상관계수를 확인하여 타겟 컬럼과 연관성이 낮은 변수를 제거해 변수를 선택하였다. 사고 심각도와의 상관계수를 바탕으로 비교적 사고 심각도와 관련이 없는 구, 일, 월 컬럼을 삭제하였다. 위와 같은 처리를 거친 변수 항목은 다음과 같다.



수치형 변수	sinmonth, cosmonth, sindate, cosdate, 평균속도
범주형 변수	법규위반, 가해운전자차종, 가해운전자성별, 피해운전자차종, 피해운전자연령대, 공휴일, 주말여부, 사고유형 2, 도로형태 2, 군집, 기상노면상태
타겟	사고심각도

- 분석의 틀

훈련 데이터와 학습 데이터는 선행연구를 참조하여 비율을 설정하였다. [9] 사이킷런의 train_test_split() 모듈을 사용하여 연도별 데이터를 골고루 섞은 뒤 전체 데이터의 80%를 학습 데이터로, 나머지 20%의 데이터를 테스트 데이터로 생성하였다. 현재 사용하고 있는 교통사고 데이터는 데이터 특성상 타겟컬럼 중 중사상사고의 비율이 적기에 타겟 컬럼의 비율을 보존하도록 나누었다.

사용 데이터는 많은 범주형 변수들을 가지고 있기에 이를 인코딩 처리해야 한다. LightGBM 과 CatBoost 는 모델이 자체적으로 범주형 변수들을 처리할 수 있기에 원핫인코딩을 하지 않고 사이킷런의 resample 함수를 활용하여 업샘플링만 수행하여 따로 전처리하였다. 이외의 모델은 범주형 변수를 학습할 수 있도록 사이킷런 preprocessing 모듈의 원핫인코더를 활용하여 원핫인코딩을 하였다. 수치형 변수들은 값의 범위에 따라 영향을 받는 것을 막기 위해 StandardScaler 를 활용하여 데이터의 평균을 0, 표준편차를 1 로 변환하는 표준화를 진행하였다. 그후, 앞의 두 모델과 같은 방식으로 업샘플링을 진행하였다.

- 모델 선택

1. Random Forest

랜덤 포레스트는 앙상블(Ensemble) 학습을 이용한 기계학습 알고리즘이다. 앙상블은 여러 분류기를 생성한 뒤 결합하여 그 분류 모델들에서 나온 결과들을 투표나 가중 평균을 통해 최종 예측을 하는 기법이다. RF 는 배깅(bootstrap Aggregating)을 기반으로 하여 중복을 허용한 서브셋 샘플을 생성하며 각 결정 트리마다 다른 데이터 샘플로 학습을 한다. 랜덤 포레스트는 결정 트리를 사용하여 빠른 학습 속도를 가지고 많은 feature 을 처리할 수 있으며 다양한 분류기를 통해 과적합을 방지한다.

2. Extra Trees

Extra Trees 는 Random Forest 의 변형 모델이다. RF 의 경우 최적의 분할을 찾기 위해 Information Gain 을 계산하지만 Extra Trees 는 각 후보 특성을 무작위로 분할하여 무작위성을 더욱 증가시킨다. 또 훈련 시 Bootstrap 샘플을 생성하지 않고 전체 데이터 샘플을 사용한다. 이러한 방법으로 Random Forest 보다도 연산 속도가 빠르고 Bias 와 Variance 를 낮출 수 있다는 장점이 있다.

3. XGBoost

XGBoost 는 부스팅 알고리즘을 기반의 모델이다. 부스팅은 이전 분류기의 학습 결과를 토대로 다음 분류기의 학 습 데이터의 샘플 가중치를 조정해 학습을 진행하는 방법이다. 부스팅 방식은 Bootstrap 샘플을 생성하고 분류기를 학습시킨다. 그 결과를 이용해 잘못 분류된 데이터와 학습에 사용되지 않은 데이터에는 가중치를 부여하여 다음 학습을 하여 오차를 보완한다. 이러한 과정을 반복해 최종 모델을 생성한다. XGBoost 는 greedy 알고리즘이나

approximate 알고리즘 등을 사용해 Loss function 이 최소화되도록 하는 최적의 Split point 를 찾는다. 이는 큰 데이터에서도 높은 분류 성능을 보인다. [14]

4. LightGBM

LightGBM 알고리즘은 XGBoost 의 단점을 개선한 방법이다. 속도와 Overfitting 을 방지하기 위해 만들어졌으며 결정 트리 기반으로 리프 중심의 트리 분할(Leaf-wise tree growth)이라고 불리는 방식을 사용한다. 이는 균형적으로 트리를 분할하기보다 손실을 최대한 줄일 수 있는 리프를 우선적으로 분할하며 이는 빠른 학습 속도와 메모리 사용량을 적게 사용하면서도 높은 정확도를 보인다. 또 큰 규모의 데이터에서 효율적으로 사용할 수 있다. [9]

5. CatBoost

CatBoost 범주형 변수 데이터 처리에 유용한 GBM 기반의 알고리즘이다. Ordered Boosting 기법을 활용해 범주형 변수의 전처리와 과적합 문제를 보완한다. Ordered Boosting 은 일부 데이터의 잔여 오차를 이용하여 모델을 생성하고, 다음 데이터의 잔차를 이용하여 모델이 예측한 값을 다시 사용한다. 또 Ordered Target Statistics 방법을 사용한다. 이는 훈련 데이터에 순서를 부여하고, 각 범주값의 타깃 평균을 계산하는데 현재 데이터는 제외하고 훈련된 데이터까지만 계산한다. 이렇게 현재 데이터를 인코딩하여 타깃 유출과 과적합을 방지한다. [15]

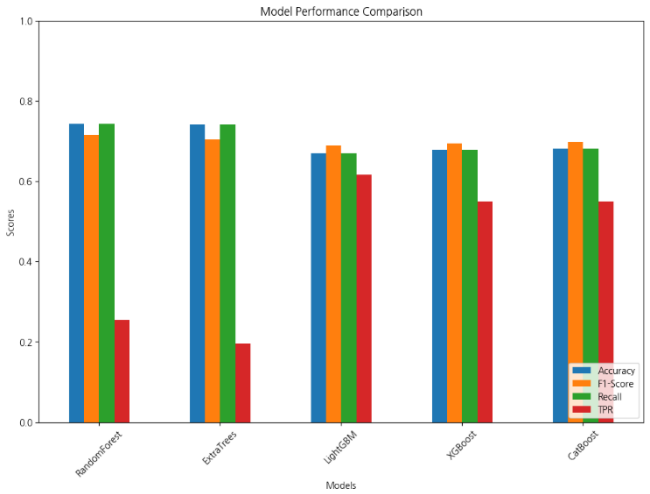
Catboost 알고리즘은 다른 머신러닝 모델들과 달리 범주형 변수에 대한 인코딩이나 변수 선택 없이 사용할 수 있다. 또 초기 하이퍼 파라미터값이 최적화된 상태로 별도의 파라미터 튜닝을 하지 않아도 된다는 장점이 있다.

III. Result

- 실증 분석

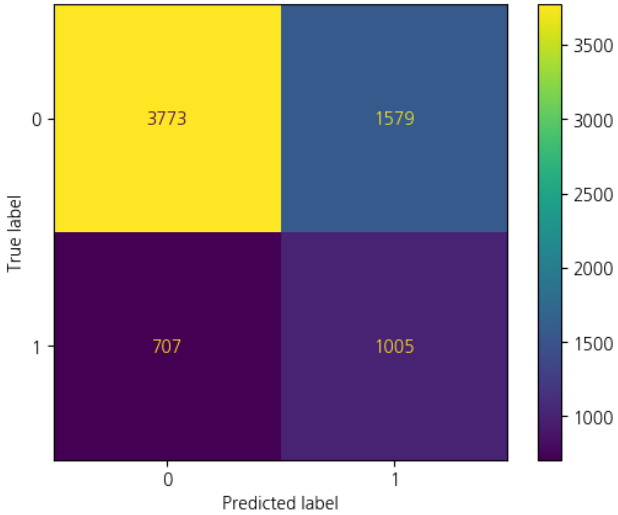
모형의 예측력 평가는 데이터 셋의 불균형 함을 고려하고자 confusion matrix 를 기반으로 예측 성능 지표로 Accuracy, Recall, F1 score, TPR 를 이용하였다.

F1 score 를 기반으로 보면 각 모델의 전반적인 예측력은 0.7 내외로 큰 차이를 보이지 않는다. 하지만 Bagging 기법을 적용한 모델들이 accuracy 가 높고 TPR 이 낮은 반면 Boosting 기법을 적용한 모델들이 accuracy 가 낮은 대신 TPR 이 더 높은 차이를 보인다. 이는 중상, 사망사고로 분류되는 사고 심각도가 경, 부상보다 훨씬 적기 때문에 Bagging 모델들이 중사상 사고를 예측하는데 한계를 보임을 알 수 있다. 이를 개선하기 위해 Randomforest 에 차원 축소 기법 중 하나인 UMAP 을 적용해 본 결과 타겟값에 대한 recall 은 0.25 에서 0.44 로 개선된 반면 f1 스코어는 0.72 에서 0.68 로 감소했고 때문에 최종 모델로 선택하지 않았다. 반면 Boosting 모델들은 accuracy 가 낮은 대신 타겟 값에 대해 더 정확한 예측을 하고 있다. 도로교통공단에 따르면 사망사고의 비용은 5 억 3379 만원으로 경상 520 만원에 비해 100 배 이상이며 중상 사고 역시 6,890 만원으로 경상사고의 10 배 이상이다. [16] 그런 만큼 교통사고 심각도에서 중사상 사고를 예측하는 것이 중요하기 때문에 Boosting 모델을 선정했고 그 중 recall(0.68)이 가장 높은 Catboost 를 최종적으로 최적의 모델로 선정해 TPR 을 높이는 방향으로 파라미터를 최적화했다.



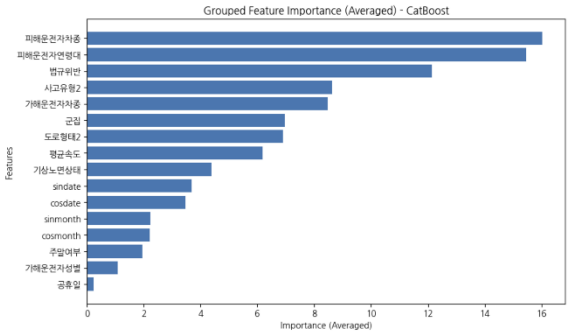
- 파라미터 최적화

위에서 설명한 바와 같이 Catboost 는 자체적으로 파라미터를 튜닝하지만 TPRl 값이 중요하다고 생각한 바 이를 높이기 위해 튜닝을 진행했다. [17] 규제 적용을 위해 max_depth 를 1 에서 5 사이와 min_child_sample 1 에서 40 사이로 조정했고 그 결과 최적의 파라미터는 params = {'depth': 8, 'grow_policy': 'SymmetricTree', 'learning_rate': 0.015930522616241012, 'min_child_samples': 24}로 나왔다. 그 결과 TPR 값은 0.55 에서 0.58 로 개선되었다.



- 특성 중요도 분석

고령운전자의 사고 심각도 판단에 있어 중요한 특성을 확인하기 위하여 최종 모델로 선택된 CatBoost 모델의 get_feature_importance 기능을 활용하였다. 이를 통해 도출된 최종 모델의 특성 중요도는 다음과 같다.



교통사고 심각도의 중요한 영향 요인은 피해운전자 차종, 피해운전자 연령대, 법규위반, 사고유형 2, 가해운전자 차종 순으로 나타났다. 이중 상위 3 개의 특성에 대해 어떠한 영향을 미치는지 더 자세히 분석하고자 SHAP 분석을 수행하였다. 먼저, 피해운전자 차종에 대한 SHAP 분석을 진행하였다. 이륜, 자전거, 원동기가 높은 SHAP 값을 보이며 중사상 사고에서 중요한 기여 요인으로 나타났다. 해당 차종들이 사고가 발생할 경우 중사상 사고로 이어질 가능성을 더 많이 높였다는 것을 의미한다. 다음으로는, 피해운전자 연령대 컬럼에 대해 SHAP 값을 분석했다. 65 세 이상, 51-60 세, 61-64 세 순으로 중사상 사고 발생에 크게 기여하였다. 주로, 피해자가 고령일수록 사고가 발생할 경우 중사상 사고로 이어질 가능성이 상대적으로 높다는 점을 시사한다. [18] 마지막으로, 법규위반 특성에 대해 SHAP 분석을 진행하였다. 과속이 0.82 로 가장 높은 값을 가지며 중사상 사고 발생 가능성에 큰 영향을 미쳤다. 이외에도 신호위반, 중앙선 침범, 보행자 보호 의무 위반 순으로 중사상 사고가 발생할 가능성을 높이는 데에 기여하였다.

IV. Conclusion

본 프로젝트는 2018년부터 2023년까지의 서울시 65 세 이상의 고령 운전자들의 교통사고 데이터 자료와 기계학습을 활용하여 고령 운전자 교통사고 심각도 예측을 수행하였다. 총 5가지 모델을 활용하여 분석을 하였으며, 종합적으로 보았을 때 CatBoost가 최종 모델로 선택되었다. 기존의 선행연구는 비교적 과거의 데이터를 활용하여 주로 전체 운전자 집단에 대한 분석을 하였다면, 최근 중요한 사회적 논의 대상인 고령 운전자를 대상으로 최신의 데이터를 활용하여 분석하였다는 점에서 의의가 있다. 또한 단순 사고 심각도를 분류하는 것에 그치지 않고, 특성 중요도를 활용하여 고령 운전자의 사고 심각도에 영향 요인을 분석하고, 각 특성에서 중요한 영향을 미치는 범주들을 확인하였다. 고령운전자의 사고 심각도 분류 모델을 구축하여 위험도가 높은 중사상사고에 대해 예측을 할 수 있다. 더불어, 특성 중요도를 이용하여 정책적 제안에 활용할 수 있다. 먼저, 사고 심각도에 가장 큰 영향을 미친 피해운전자 차종 중에서도 이륜차, 자전거, 원동기와 같은 중사상사고 취약 차종의 이용자들을 위한 움직임이 필요하다. 해당 차종 이용자들에게 충돌 방지 센서를 설치하거나 보호 장구 착용 단속을 더 강하게 하여 중사상사고를 예방할 필요가 있다. 또한 피해운전자도 마찬가지로 고령일수록 중사상사고로 이어질 가능성이 더 컸기에, 그들을 위해 주기적인 교통안전교육을 시행하는 등 안전교육을 확대하고 지속적인 홍보를 할 필요가 있다. 또한 고령운전자의 시계 향상을 위하여 교통 표지판의 글자 크기를 확대하는 등의 고령운전자를 위한 교통 인프라를 개선할 필요가 있다. 마지막으로 법규 위반 중 중사상사고로 이어질 위험성이 큰 과속, 신호위반, 중앙선침범에 대해 단속을 특히 강화할 필요가 있다. 과속방지 시설을 확대하고 이동식 단속 카메라를 설치하여 이러한 법규에 대한 위반을 줄일수록 있다고 노력해야 한다.

다만, 본 프로젝트는 여러 한계점으로 인해 높은 분류 성능을 보이지는 못했다. 먼저, TAAS 교통사고분석사고시스템으로부터 얻은 실세계 공공데이터를 활용하였는데, 이러한 데이터에는 범주형 데이터가 대부분이었다. 수치형 데이터보다는 주로 범주형 데이터에 대한 정보가 많았기에 이를 활용하여 분류 성능을 높이는 데에 한계가 존재했다. 범주형 변수에서 파생변수를 추출하고, 월별 구별 평균속도와 같은 추가적인 특성을 도입하여 수치형 변수를 활용하고자 하였으나 어려움이 있었다. 또한, TAAS에서 제공하는 데이터를 활용하다 보니, 사고심각도라는 타겟 컬럼에 강한 영향을 미치는 의미 있는 특성들이 많이 존재하지 않았다. 다양한 방식으로 특성을 추가하였으나 역시나 한계가 존재하였다. 안전벨트 착용 여부, 차량 내에서의 운전자 자리, 차량 무게 등의 사고 심각도와 더 강한 관계가 있을 만한 특성을 제공받지 못했다는 점에서 아쉬움이 있다. 앞서 추가한 평균 속도 또한 일별 데이터를 확보에 어려움을 겪어 월별 데이터를 추가했기에 이를 보완할 필요가 있다. 더불어, 선행연구에서도 언급되었듯 교통사고 데이터의 고질적인 문제인 데이터 불균형은 피해갈 수가 없었다. 이를 해결하기 위해 업샘플링을 진행하고 가중치를 두어 학습을 시키는 등의 시도를 하였으나, 중사상사고는 경상사고에 비해 빈도가 낮을 수밖에 없기에 이를 완전히 해결하지는 못하였다. 후속 연구에서는 이러한 데이터의 특성과 다양한 의미 있는 특성들의 도입을 통해 분류 성능을 높일 다양한 방법에 대한 추가적인 연구가 필요하다.

V. References

[1] 도로교통공단. (2023). “전체 교통사고는 줄어드는데 고령운전자 교통사고는 증가”, 2023.10.26., https://www.koroad.or.kr

[2] 정시내. (2024). “지난해 5 건 중 1 건은 고령운전자 사고…"정부 대책 강구해야"”. 중앙일보, 2024.9.30. https://www.joongang.co.kr/article/25281045.

[3] 윤보람. (2024). “68 세 운전자 역주행 사고로 9 명 사망…자격 논란 재점화되나”. 연합뉴스. 2024.7.2. https://www.yna.co.kr/view/AKR20240702002700004.

[4] 정윤주. (2024). “"65 세 이상 운전자 면허 자진반납률, 부산·서울이 비교적 높아”. 연합뉴스. 2024.09.18 . https://www.yna.co.kr/view/AKR20240913082700004.

[5] 박세영, 송영훈, 김광오, 한결아,& 조석헌. (2024). 기계학습 알고리즘을 이용한 교통사고 심각도 예측 모델에 관한 연구. 한국통신학회 하계종합학술발표회 논문집, pp. 1211-1214.

[6] 김남현, 고상근, 김성재,& 이수안. (2022). 도로 형태와 지역 정보를 결합한 서울시 교통사고 건수 예측 모델. 한국정보과학회 2022 한국소프트웨어종합학술대회 논문집, pp. 1382-1384.

[7] TaeWook Kim, JiWoong Yang, Hyeonjin Jung, HanJin Lee and Ellen J, Hong(2024), “Group Clustering Technique and Risk Estimation Method for Traffic Accident Prevention”, Journal of The Korea Society of Computer and Information, Vol. 29 No. 8(2024) pp.53~58

[8] 남명우, 박두서, 장영준,& 이홍철. (2021). 머신러닝을 이용한 교통사고 사상자 수 예측: 서울시 공공데이터를 대상으로, 한국컴퓨터정보학회 동계학술대회 논문집, 29(1), pp.27-30

[9] 이현미, 전교석, & 장정아. (2020). LightGBM 알고리즘을 활용한 고속도로 교통사고 심각도 예측모델 구축. 한국전자통신학회 논문지, 15(6), pp. 1123-1130.

[10] 이혜령, 금기정,& 손승녀. (2011). 고속도로 교통사고 심각도 등급별 요인분석에 관한 연구. 한국도로학회논문집, 13(3), pp. 157-165.

[11] Taas.koroad.or.kr. 2020. TAAS 교통사고분석시스템

[12] Seunghoon Kim(2022), “Elderly Driver-involved Crash Analysis and Crash Data Policy”, Korean Inst.Intelligent.Transp.syst, Vol. 21 No.5(2022) pp.90~102

[13] 김승훈, 임영빈, & 김기정. (2021). 머신러닝 기반의 수도권 지역 고령운전자 차대사람 사고심각도 분류 연구. 디지털융복합연구, 19(4), pp. 25-31.

[14] 정자훈. "XGBoost 를 활용한 EBM 3D 프린터의 결함 예측." Journal of the Korea Institute of Information & Communication Engineering 26.5 (2022).

[15] 소욱, 이다인, & 공민정. (2023). 청년층 초기 일자리의 시퀀스 특성이 경력정착기 노동시장 성과에 미치는 영향: OMA 와 CatBoost 기법을 중심으로. 직업능력개발연구, 26(3), pp. 135-171.

[16] 도로교통공단. (2024). “2022 년 도로교통사고 사회적 비용, 약 26 조 2,833 억 원 발생”, 2024.2.15., https://www.koroad.or.kr

[17] 이시욱. (2024). 노드 크기와 깊이에 따른 앙상블의 성능과의 관계. 석사학위논문, 고려대학교 대학원 통계학과, 서울.

[18] Kim, T. S., Lee, K. H, King, T. H. KIM, O. H., Cha, Y. S., Cha, K. C. and Hwang, S. O.(2014), “Clinical Characteristics and Prognostic Factors of Geriatric Patients Involved in Traffic Accidents”, Journal of Civil and Environmental Engineering Research, vol. 34, no. 4, pp.1279~1287