Regular paper | Journal of the KIECS. pp. 1123-1130, vol. 15, no. 6, Dec. 31. 2020, t. 104, pISSN 1975-8170 | eISSN 2288-2189
http://dx.doi.org/10.13067/JKIECS.2020.15.6.1123

LightGBM 알고리즘을 활용한 고속도로 교통사고심각도 예측모델 구축

이현미* · 전교석** · 장정아***

Predicting of the Severity of Car Traffic Accidents on a Highway Using Light Gradient Boosting Model

Hyun-Mi Lee* · Gyo-Seok Jeon** · Jeong-Ah Jang***

요 약

본 연구는 고속도로 교통사고 심각도 예측모델을 구축하기 위해 다섯가지 머신러닝 기반의 분류모형 적용하였다. 2015년~2017년 동안 전국 고속도로에서 발생한 사고 데이터 21,013건을 5가지의 분류 모형을 적용한 결과 LightGBM(Light Gradient Boosting Model)이 가장 좋은 성능을 나타내는 것으로 나타났다. LightGBM에서는 교통사고심각도 추정에 있어 우선순위 요인으로 사고차량 수, 사고유형, 사고지점, 사고차로유형, 사고차량 유형 순으로 나타났다. 이러한 모형의 결과를 기반으로 일관적인 사고심각도 예측 과정을 통하여 교통사고대응관리 전략 수립에 활용할 수 있다. 본 연구는 국내 기계학습을 활용한 사례가 적은 여건에서 향후 빅데이터 기반의 다양한 기계학습 기법을 활용이 가능함을 제시하고 있다.

ABSTRACT

This study aims to classify the severity in car crashes using five classification learning models. The dataset used in this study contains 21,013 vehicle crashes, obtained from Korea Expressway Corporation, between the year of 2015 - 2017 and the LightGBM(Light Gradient Boosting Model) performed well with the highest accuracy. LightGBM, the number of involved vehicles, type of accident, incident location, incident lane type, types of accidents, types of vehicles involved in accidents were shown as priority factors. Based on the results of this model, the establishment of a management strategy for response of highway traffic accident should be presented through a consistent prediction process of accident severity level. This study identifies applicability of Machine Learning Models for Predicting of the Severity of Car Traffic Accidents on a Highway and suggests that various machine learning techniques based on big data that can be used in the future.

키워드

Accident Severity, Influencing Factor, LightGBM, Machine Learning, Prediction Model,
사고 심각도, 영향 요인, 라이트 그레디언 부스팅 모형, 기계 학습, 사고 등급 예측 모형

* 아주대학교 교통시스템공학과 박사과정(hm0625@ajou.ac.kr)
** 제2저자 : 아주대학교 교통시스템공학과 연구교수 (wjsrytjr@ajou.ac.kr)
*** 교신저자 : 아주대학교 교통시스템공학과 연구교수
· 접 수 일 : 2020. 10. 07
· 수정완료일 : 2020. 11. 11
· 게재확정일 : 2020. 12. 15
· Received : Oct. 07, 2020, Revised : Nov. 11, 2020, Accepted : Dec. 15, 2020
· Corresponding Author : Jeong-Ah Jang
TOD-based Sustainable City Transportation Research Center, Ajou University,
Email : azang@ajou.ac.kr

Ⅰ. 서 론

2017년 기준 우리나라 교통사고는 총 114만 3천 175건으로 그 피해액을 경제적 가치로 환산하면 약 40조 574억원[1]으로 한국의 연간 GDP의 2.3%에 해당하는 규모이다. 특히 고속도로 교통사고는 전체 교통사고 건수의 2%에 불과하지만 사망자수는 전체의 6%, 부상자수는 3%에 달하는 등 사고건수는 적지만 동일한 사고가 나더라도 심각한 사고가 발생하고 있다. 또한 고속도로의 2차 교통사고로 인한 치사율은 약 50%로 전체 교통사고의 치사율보다 3.3배 정도 높은 등 안전도상 피해 규모가 크고 도로망에 미치는 정체 파급효과의 강도가 높고, 넓은 범위에 영향을 미친다.

이러한 측면에서 고속도로 사고 발생 시 신속하고 효과적인 방법으로 대처하는 것이 중요하다. 신속하게 사고처리를 수행하여 사고 발생 지점의 통행을 신속히 정상화하는 것과 적합한 범위의 고속도로 이용자들에게 정보를 제공하여 대안 경로 등을 통한 교통사고 발생 후 대응 전략이 필요하다.

일반적으로 교통사고발생에 따른 대응전략에서 그 범위를 설정하는 문제에서 사고처리 시간이 주요한 변수가 된다. 사고처리 시간에는 사고가 야기한 손실 크기, 즉 사고심각도 및 교통사고등급이 밀접하게 관련되어 있다. 즉, 교통사고 발생 시 교통사고등급을 빠르게 예측하여 사고처리를 위한 정보를 제공 및 지원하는 것이 중요하다[1].

현재 우리나라 고속도로 교통사고 데이터의 경우 한국도로공사에서는 자체기준으로 사고피해액, 사상자, 관련 차량 수를 기준으로 사고의 피해 정도에 따라 교통사고 심각도를 등급화하여 A-D급으로 분류하고 있다. 그러나, A-D 등급이라는 심각도 등급은 관리자에 따라 다른 등급으로 분류되어 일관성이 결여되는 문제[2]가 지적된 바 있으며, 사고처리 이후 중상자의 사망과 같은 내용 갱신에 대한 자세한 자료는 부족한 실정이다. 현재 교통사고등급 자료는 기본적으로 일관성과 부정확성 문제가 내재되어 있으며, 자료 특성 상 이상치나 부정확한 자료를 선별하기 어렵다는 특성을 가지고 있다.

본 연구는 고속도로 교통사고 대응 및 사고로 인한 비반복적 정체 대응 측면에서 사고처리시간의 정확한 예측이 요구된다는 점과 이를 위해서는 사고발생 시 빠르고 일관성 있는 교통사고등급 분류 및 예측 방법론이 필요하다는 측면에서 연구를 수행하였다. 빅데이터 분석에 많이 적용되는 기계학습기법 중 하나인 LightGBM(Light Gradient Boosting Model, 이하 GBM) 알고리즘을 활용하여 고속도로 교통사고 발생 시 사고심각도를 빠르고 일관성 있게 예측하고자 하였다. 연구결과는 사고 발생 시 신속한 교통사고심각도등급 예측과 함께 기존 교통사고등급 자료 상 부정확한 자료의 선별, 이를 통한 자료 정확성 및 관련 모형의 정확성 제고에 기여할 수 있을 것으로 기대된다.

본 연구는 고속도로 교통사고 심각도 등급을 예측하고 영향요인의 우선순위를 도출하기 위하여 전국 고속도로 교통사고 21,013건을 분석 대상으로 하였다. 먼저 본 논문의 제2장에서 국내외 연구 검토를 통해 사고심각도등급 예측에 적용된 다양한 방법론을 검토하였다. 제3장에서 국내 고속도로 데이터 분석을 위해 연구 범위와 방법론을 설정하였고, 제4장에서 사고심각도등급 예측에 적합한 기계학습 모형을 제시하기 위해 그림 1과 같은 분석 방법론을 제시하였다. 제5장에서는 선택된 최적의 모형과 기존 통계적 모형 간의 정확도를 비교하였으며, 마지막으로 제6장에서 연구의 결과와 한계점을 서술하였다.

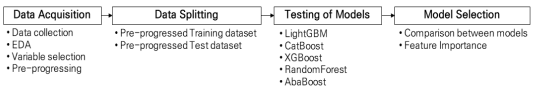

그림 1. 분석 방법론 및 연구 프로세스
Fig. 1 Visual representation of analysis process

Ⅱ. 선행연구 검토 및 차별성

기존 국내외 연구에서는 교통사고 심각도를 예측하고 영향요인을 도출하기 위해 통계적 분석방법과 기계학습 접근 방식이 적용되어왔다.

통계적 분석방법은 이분형 로지스틱, 순서형 로짓 또는 프로빗 모형, 다항 로짓 회귀모형이 주로 적용되

1) https://www.koti.re.kr/user/bbs/BD_selectBbs.do?q_bbsCode=1005&q_bbscttSn=20190723133356319

었으며, 여러 가지 설명 변수를 선택적으로 투입하거나 유의한 변수를 추출하여 사고심각도에 미치는 영향에 대해 설명하는 방식에 중점을 두고 있다. 전반적으로 이러한 통계적 방법은 대부분 추론 및 추정을 위해 설계되었으며 모델 개발 전에 몇 가지 가설과 제한 사항을 충족해야한다[3]. 이는 엄격한 수학적 기초를 기반으로 예측하므로 추정기와 설명 변수 사이의 관계를 해석하여 현상을 설명 할 수 있지만 종속변수에 영향을 미치는 많은 변수들 사이의 복잡한 상호작용을 반영하지 못할 수 있다.

이에 비해 기계학습의 경우 통계적 방법과 달리 해석보다는 예측을 목표로 하며 종속 변수와 독립 변수 사이의 복잡하고 높은 비선형 관계를 처리할 수 있으며[4] 통계적 모형보다 높은 예측 정확도를 보이는 것으로 나타났다[5]. 국외 연구에서는 교통사고심각도 예측을 위해 다양한 기계학습 기법들의 적용 및 비교 연구가 진행되어왔으나[6, 7, 8, 9], 국내 연구에서는 고속도로 사고를 대상으로 교통사고심각도 예측을 위해 기계학습 기법을 적용한 연구는 상대적으로 부족하고 적용된 모형도 제한적인 상황이다[10].

본 논문은 학습데이터를 기반으로 부스팅(Boosting) 계열의 네 가지 기계학습 알고리즘과 그 외 알고리즘을 포함하여 총 5개 알고리즘을 대상으로 기계학습모형을 적용하였다. 그 중 가장 우수한 LightGBM 알고리즘을 활용하여 교통사고심각도등급 예측 모형에 대하여 상세히 살펴보았다. 그 동안 국내의 기계학습 적용 연구가 부족한 바 관련 논의를 보다 풍부하게 할 수 있다는 점에서 본 연구의 의의가 있다. 그리고 LightGBM의 경우 영향요인의 중요도를 제시하므로 해석적 측면에서 유용하며, 관련 논의가 가능하다는 데에도 차별성을 지닌다.

III. 방법론 및 연구 고찰

3.1 Gradient Boosting Model

Gradient Boosting Model은 서로 다른 개별 모형들을 결합하여 모형의 성능을 높이는 앙상블 기법 중 하나인 부스팅(Boosting) 계열의 알고리즘이다. 부스팅 기법은 약한 분류기인 의사결정나무모형 여러 개를 순차적으로 적합하고 이를 하나로 묶어 만든 모형으로 예측한다. 첫 번째 구축된 나무모형에서의 예측결과는 원래 데이터와 비교하여 잔차를 생성하는데, 생성된 잔차는 두 번째 나무모형의 대상이 된다. 고정된 임계값 아래의 잔차가 나타날 때까지 많은 나무를 구축하고 모든 나무의 결과를 병합하여 모형을 제시한다. 또한 빠르고 다양한 최적화 옵션을 제공하고 자동 가지치기 기능이 있어 과적합을 피할 수 있다는 장점을 지니고 있다[13]. 이 때문에 효율성, 정확도, 해석가능성 측면에서 장점을 지니는 기계학습 알고리즘으로 알려져 있다. GBM은 식 (1)과 같이 나타낼 수 있다[14].

$$F(x)=\sum_{m=1}^{M}F_m(x)=\sum_{m=1}^{M}\beta_m h(x;a_m) \quad (1)$$

$F(x)$는 x변수들 기반으로 반응변수 y의 근사함수를 나타낸다. $h_m(x;\alpha_m)$는 α_m의 파라미터를 갖는 m개의 의사결정 나무이며 β_m은 손실함수 $L(y,F(x))=[y-F(x)]^2$의 최소화에 따라 값이 결정되는 값이다. 이때 경사하강방식으로 특정 손실 함수의 예상 값을 최소화하여 모형을 업데이트한다. 위의 논의를 바탕으로 GBM은 그림 2와 같이 요약할 수 있다.

Initialize $F_0(x)$ to be a constant, $F_0(x)=\arg\min_{\beta}\sum_{i=1}^{N}L(y_i,\beta)$

For m=1 to M:

For $i=1,2,\cdots,N$ compute the negative gradient

$$\tilde{y}_i=-\left[\frac{\partial L(y_i,F(x_i))}{\partial F(x_i)}\right]_{F_m(x)=F_{m-1}(x)}$$

Fit a decision tree $h(x;a_m)$ to the targets $\tilde{y}_i$ given terminal regions $R_{jm}, j=1,2,\cdots,J_m$

Compute a gradient descent step size as

$$\beta_m=\arg\min_{\beta}\sum_{i=1}^{N}L(y_i,F_{m-1}(x)+\beta h(x;a_m))$$

Update the model as $F_m(x)=F_{m-1}(x)+\beta_m h(x;a_m)$

Output the final model $F(x)=\sum_{m=1}^{M}F_m(x)$

그림 2. 그래디언 부스팅 모형 알고리즘
Fig. 2 Algorithm for the gradient boosting decision trees

3.2 Light Gradient Boosting Model

LightGBM 알고리즘은 의사 결정 트리 알고리즘에 기반한 고성능의 알고리즘으로 순위 또는 분류를 위한 기계학습 작업에 사용되고 있다. GBM이 높은 예측력으로 다양한 분석에서 사용되었지만 고차원 변수가 포함된 빅데이터에 적용 시 훈련 속도와 메모리 소비면에서 비효율적이라는 단점을 가지고 있다. 이를 해결하기 위해 2017년 Microsoft에서 Gradient-based One-Side Sampling(GOSS)을 통해 데이터의 일부만

으로 빠르게 정보이득을 계산하고 Exclusive Feature Bundling(EFB)으로 특성요인들을 줄이는 알고리즘을 개발하여 모델링 시간을 단축시키는 방법인 LightGBM이 소개되었다[15]. GOSS 방식은 Gradient의 크기를 기반으로 샘플링하며 데이터 갯수를 내부적으로 줄여서 계산한다. 큰 Gradient를 갖는 데이터들은 모두 남겨두고, 작은 Gradient를 갖는 데이터 개체들에서는 무작위 샘플링을 진행한다. EFB방식은 희소한 변수 공간의 특성에 따라 배타적인 변수들을 하나의 변수로 통합하며 이를 통해 계산량을 줄인다.

LightGBM은 의사 결정 트리 알고리즘을 기반으로, 트리의 깊이(depth wise)나 균형 트리(level wise)로 분할하는 다른 부스팅 알고리즘과 달리, 가장 잘 맞는 트리의 리프중심(leaf wise)으로 분할한다. 따라서 LightGBM에서 동일한 리프(leaf)에서 성장할 때 리프(leaf)방식 알고리즘은 균형 레벨(level)방식 알고리즘보다 손실을 더 줄일 수 있으므로 기존 부스팅 알고리즘으로는 거의 달성하기 어려울 만큼의 정확도를 달성하면서도 매우 빠른 수행이 가능하다.

또한 LightGBM은 영향요인의 중요도를 제시하여, 기존 모형들보다 해석적 측면에서 유용하다고 할 수 있다.

Ⅳ. 분석 데이터 및 모형검증 및 평가방법

4.1 분석 데이터 및 요인별 현황

본 연구는 2015년부터 2017년까지 전국 고속도로에서 발생한 사고 데이터 21,013건을 분석에 활용하였다. 고속도로 교통사고 데이터는 사고 발생 정보(사고발생 요일, 사고 발생 시간, 날씨정보), 사고지점의 도로 시설 유형(사고발생지점, 사고발생지점 상세, 사고차로유형, 방책시설 유형), 사고지점 도로특성(도로종단경사, 노면상태) 사고특성(사고유형, 전복사고, 화재사고, 잭나이프사고, 언더라이드사고), 교통위험특성(교통장애요인, 도로장애요인) 그 외(사고차량 유형, 관련사고차량 수) 등의 항목을 포함하고 있다. 이와 관련된 주요 기술 통계는 표 1과 같다.

표 1. 분석데이터 기술통계
Table 1. Descriptive statistics for selected variables for analysis.

Category	Variable	Total	Number of missing	Description
Environment Characteristics	Time of Day	21,013	0	Night (8PM-5AM); Day (6AM-8PM)
	Weather Condition	21,001	12	Severe Crosswinds; Snow; Fog; Rain; Hail; Blowing Sand; Sunny
Facility Characteristics	Incident Point	21,013	0	TollGate; Ramp; Road Work Zone; Station; Tunnel
	Incident Point2	20,973	40	Link segment, Highway shoulder; Diamond-shaped Interchanges; Clover-shaped Interchanges, Trumpet Interchanges, Entrance Ramp; Others40
	Incident line type	13,293	7,720	Reversible lane; Acceleration Lane; Deceleration Lane; Exclusive Bus Lane; Shoulder; Road Work Zone; Entrance of highway; Exit of highway; Intersections
	Facility (Median)	20,654	359	Guardrail; Barrier(Fixed Barrier, Stiffened Barrier, Concrete Barrier, Mobile Barrier); No Median; Others
Road Characteristics	End Slope	21,013	0	Flat; Slope; Downhill
	Surface Status	21,013	0	Dry; Humidity; Snow; Others
Accident Characteristics	Type of Accident	20,569	444	Jaywalking-pedestrian crash; Frontal crash; Rear-end crash; Side-on crash, Others
	Rollover	21,013	0	Rollover
	Fire	21,013	0	Fire
	Jacknife	21,013	0	Jacknife
	Underride	21,013	0	Underride
Risk Characteristics	Traffic Problem	20,936	77	Stationary vehicle on the highway shoulder; Traffic congestion; Unexpected situation congestion; congestion caused by work; No other problem
	Obstructive Factor	20,936	77	Fallen object; Water-Reservoir; Slippery surface; Deformation; Soil erosion; Pot hall; Others
Others	Vehicle type	19,171	1,842	Compact Vehicle; Subcompact Vehicle; Midsize vehicle; Sport utility vehicle; Heavy Goods Vehicles(Bus); Truck; Tank lorry; Others
	Accident involvement	21,013	0	Number of involved cars in road accidents

4.2 학습 및 평가 데이터 분리

예측 모형의 과대 추정(over-fitting)의 문제를 피하기 위해 전체 데이터를 고속도로 교통사고 예측모형 구축을 위한 학습데이터와 예측력을 검증하는 테스트 데이터로 구분하였다. 모형구축을 위한 학습 데이터로 전체(21,013건)의 80%인 16,810건 데이터를 무작위 추출하고, 20%인 4,203건을 모형 예측 및 성능 비교를 위한 모형의 '최종 성능'을 평가 테스트 데이터로 간주한다. 이후 표 2와 같이 학습 데이터와 테스트 데이터의 분포가 유사함을 확인하였다. 고속도로에서 발생하는 교통사고는 고속도로 관리기관인 한국도로공사에서 관리를 담당하며, 교통사고등급은 사고 피해 정도에 따라 크게 4가지(A-D)등급으로 분류하고 있다. 여기서, C등급 이상의 사고는 인적·물적피해가 발생한 사고를 나타내며 A등급의 피해규모가 가장 크다.

표 2. 학습 데이터와 테스트 데이터의 분포
Table 2. Division of Incident risk ranking

Rank of incident	Description	Train data	Test data	Overall
A (Coding=4)	• Over 3 fatalities • Over 20 injuries • Property damage(over ₩10,000,000)	19 (0.11%)	8 (0.19%)	27 (0.13%)
B (Coding=3)	• Over one fatality • Over 5 injuries • Property damage(over ₩2.500,000)	481 (2.86%)	130 (3.09%)	611 (2.91%)
C (Coding=2)	• Over 1 injury • Property damage(over ₩300,000)	4,274 (25.43%)	1,047 (24.91%)	5,321 (25.32%)
D (Coding=1)	• Only property damage	12,036 (71.60%)	3,018 (71.81%)	15,054 (71.64%)
Overall		16,810 (80%)	4,203 (20%)	21,013 (100%)

Ⅴ. 교통사고 심각도 예측모형 개발 및 평가

5.1 모형 평가 방법

기계학습 분류 기반으로 한 예측모형의 평가 및 성능 비교에는 정밀도(Precision), 재현율(Recall), 정확도(Accuracy), F-1 점수(F-1 Score) 등 4가지 지표가 사용된다. 이 지표들은 분류 모형의 성능 지표로서 모형은 예측결과와 실제 값을 기반으로 표 3과 같이 혼동행렬로 요약되며 4가지 지표가 산출된다.[11, 12]

- 정밀도(Precision) : 예측을 Positive로 한 대상 중 예측과 실제 값이 Positive로 일치한 데이터의 비율=TP/Pred P
- 재현율(Recall) : 실제 값이 Positive인 대상 중에 예측과 실제 값이 Positive로 일치한 데이터의 비율=TP/Act T
- 정확도(Accuracy) : 전체 예측 데이터에 대한 정확한 예측 수의 비율=(TP+TN)/All
- F-1점수(F-1 Score) : 정밀도(Precision)와 재현율(Recall)의 조화 평균==2·Precision·Recall/(Precision+Recall)

정밀도(Precision)는 예측모형으로 분류된 사고등급과 실제 사고등급과 일치한 비율을 나타내며 정밀도가 높을수록 좋은 모형이다. 재현율(Recall)은 실제 사고등급별 예측모형으로 분류된 사고등급과 실제 사고등급이 일치한 비율을 나타내며 재현율이 높을수록 좋은 모형이다. 최적모형 구축을 위해 모형의 파라미터 조정 시 알고리즘의 정밀도와 재현율은 반비례 관계를 가지게 된다. 따라서 두 지표의 성능 변화 전체를 고려한 F-1 점수(F-1 Score)로 전체 모형의 성능을 평가할 수 있다.

표 3. 혼동행력과 4가지 지표
Table 3. Confusion Matrix and Four measures

Accident Severity Level		Predicted Class		Total
		Negative	Positive	
Actual Class	Negative	TN	FP	TN+FP (Act F)
	Positive	FN	TP	FN+TP (Act T)
Total		TN+FN (Pred N)	FP+TP (Pred P)	TN+FP+FN+TP (All)

* TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

5.2 모형구축 및 평가 비교

고속도로 사고 발생 시 사고위험등급 예측모형을 구축하기 위해 모형학습을 위한 학습데이터는 총 16,810건, 모형 성능평가를 위한 테스트데이터는 총 4,203건으로 구성되어 있으며 비슷한 분포로 추출하였다. 학습데이터를 기반으로 현재 부스팅(Boosting) 계열의 분류 알고리즘으로 각광받고 있는 네 가지 부스팅 계열(LightGBM, CatBoost, XGBoost, AdaBoost)의 모형과 대표적인 앙상블(Ensemble) 알고리즘 중 하나인 RandomForest 모형을 구축하였다. Scikit-learn 라이브러리의 GridSearchCV[2]를 이용하여 각 모형별 최적의 하이퍼 파라미터 도출과 교차검증을 통해 각 알고리즘별 최적모형을 구축하였다. 교통사고등급 표본

수가 불균형이므로 교통사고 심각도 등급별로 산출되는 정밀도(Precision)와 재현율(Recall)의 가중평균값으로 나타내었다. 각 최적모형의 성능평가지표 결과는 표 4와 같다.

표 4. 최적모형의 성능평가결과
Table 4. Evaluation of machine learning classification algorithm

Machine Learning Algorithm	Light GBM	CatBoost	XGBoost	AdaBoost	RandomForest
Accuracy	**0.7642***	0.7595	0.7559	0.6505	0.7466
F-1 Score	**0.7532***	0.7439	0.7385	0.6512	0.7309
Precision	**0.7418***	0.7303	0.7222	0.6525	0.7148
Recall	**0.7650***	0.7581	0.7556	0.6500	0.7478
* Highest value					

모형평가 결과, 모형의 정확도(Accuracy) 지표는 LightGBM(76.42%), CatBoost(75.95%), XGBoost(75.59%), RandomForest(74.66%), AdaBoost(65.05%)순으로 나타나 LightGBM이 정확도 측면에서는 가장 높았으나, CatBoost와의 차이는 근소한 것으로 나타났다. 모형의 F-1 점수(F-1 Score) 지표는 LightGBM(75.32%)가 가장 높게 나타났으며 CatBoost(74.39%), XGBoost(73.85%), RandomForest(73.09%), AdaBoost(65.12%)순으로 나타났다.

다섯 개의 기계학습 모형 성능 비교는 그림 3과 같이 LightGBM이 가장 좋은 성능을 나타냈고 그 다음으로는 CatBoost, XGBoost, RandomForest, AdaBoost 순으로 나타났다. 이에 따라 LightGBM(Accuracy: 76.42% ,F-1 score:0.7532)을 최종 모형으로 채택하였다.

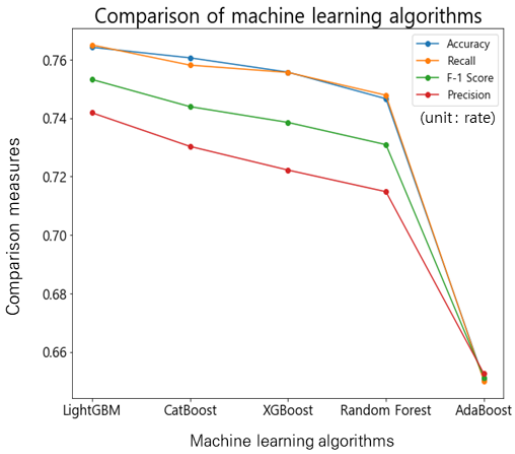

그림 3. 기계학습 모형 성능 비교
Fig. 3 Performance of candidate models

LightGBM의 사고심각도등급별 ROC 그래프 결과는 그림 4와 같이 나타났다. 최적모형의 macro AUC (Area Under the ROC Curve) 값은 0.82으로 분류기의 성능이 양호하다고 평가할 수 있다. 사고위험등급이 높은 class 3과 class 4의 AUC 값이 다른 사고위험등급들보다 높은 것으로 나타났으므로 교통사고등급이 높은 경우 분류기 예측 성능이 낮은 등급에서보다 우수하다고 해석할 수 있으며, 활용성 측면에서도 적합하다 판단된다.

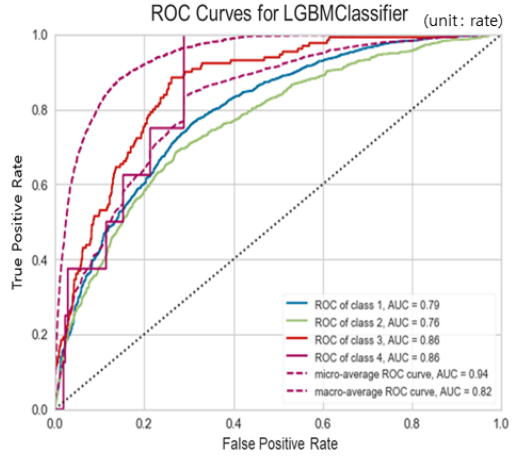

그림 4. LightGBM의 사고심각도등급별 ROC
Fig. 4 ROC Curves for LightGBM per class

2) GridSearchCV : 하이퍼 파라미터(Hyper parameter)를 순차적으로 입력하면서 최적의 파라미터를 도출하는 방안

5.3 사고심각도등급 영향요소 중요도 분석

기계학습 분류 알고리즘에 독립변수로 입력된 사고심각도 영향요소는 분류 순서, 변수에 대한 가중치 등으로 영향요인의 중요도 분석이 가능하다. 최종 채택된 LightGBM을 기반으로 교통사고등급에 영향을 주는 주요 변수를 추출한 결과는 그림 5와 같다. 최종 모형으로 선택된 LightGBM 경우, 각 변수 가중치의 절대값을 통해 변수 간 중요도를 도출하거나 줄어드는 평균 Loss값인 Gain을 통해 영향요소의 중요도를 나타낼 수 있다. 중요도 분석결과, 사고차량 수가 심각도등급을 분류하는데 가장 영향을 주는 요소로 나타났으며 그 다음으로 사고유형, 사고지점, 사고차로유형, 사고차량 유형 순으로 높게 나타났다.

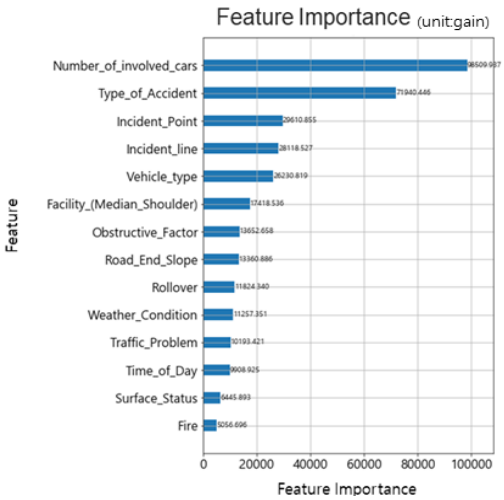

그림 5. 사고심각도등급 영향요소 중요도
Fig. 5 Feature Importance of accident severity(Gain)

Ⅵ. 결론 및 향후 연구과제

본 연구는 2015년~2017년 동안 전국 고속도로에서 발생한 사고 데이터 21,013건의 자료와 기계학습 기법을 활용하여 고속도로 교통사고심각도등급 예측 모형을 개발하였다. 학습데이터를 기반으로 다섯 가지 기계학습모형에 대해 모형별 최적 하이퍼 파라미터 도출과 교차검증을 통한 각 알고리즘별 최적 모형을 구축한 결과, LightGBM이 가장 좋은 성능을 나타내는 것으로 나타났다. 최종 모형으로 채택된 LightGBM의 교통사고등급 영향요소의 중요도는 사고차량수, 사고유형, 사고지점, 사고차로유형, 사고차량 유형 순으로 높게 나타나는 결과가 나타났다. 이러한 모형의 결과를 기반으로 일관적인 사고심각도 등급 예측 과정을 통해 교통사고대응관리 전략 수립이 제시되어야 한다.

본 연구는 국내 교통사고심각도등급 예측 모형 구축에 기계학습을 활용한 사례가 적은 여건에서 향후 빅데이터 기반의 다양한 기계학습 기법을 활용하여 사고심각도 예측을 적용할 수 있다는 측면에서 가능성을 확인하였다. 최종적으로 가장 우수한 LightGBM의 경우는 영향요인 중요도의 우선순위도도 제공하기 때문에 실제적으로 교통사고 빅데이터 활용의 필요성을 제시하여 주고 있다.

후속연구로는 현황 자료에서의 이상치, 또는 부정확 자료의 선별 기법, 이를 통한 관련 모형의 정확도 제고 효과 비교 연구 등이 필요하다. 또한 반복적 정체상황에서의 교통사고 처리전략과 비반복적 정체상황의 구분 등 빅데이터의 지속적인 기계학습 활용을 통하여 정확도 및 예측력을 향상시킬 수 있는 피드백 체계에 대한 관심과 노력이 필요할 것으로 사료된다.

감사의 글

본 연구는 국토교통부 스마트 도로조명 플랫폼 개발 및 실증 연구 개발사업의 연구비 지원(과제번호 20PQWO-B153369-02)에 의해 수행되었습니다.

References

[1] S. Lee, D. Han, and Y. Lee, “Development of Freeway Traffic Incident Clearance Time Prediction Model,” *J. of Korean Society of Transportation,* vol. 33, no. 5, 2015, pp. 497-507.

[2] J. Park, J. Jin, D. Kang, and I. Seo, “A Study on the Development of the Seasonal Highway Traffic Accident Damage Model,” *Korean Society of civil engineers*, Daejeon, South Korea, 2013, pp. 473-476.

[3] M. Karlaftis and E. Vlahogianni, “Statistical methods versus neural networks in transportation research: Differences, similarities and some insights,”

Transportation Research Part C: Emerging Technologies, vol. 19, no. 3, 2011, pp. 387-399.

[4] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote micro wave sensor data," *Transportation Research Part C: Emerging Technologies,* vol. 54, 2015, pp. 187 - 197.

[5] S. Piri, D. Delen, T. Liu, and H. Zolbanin, "A data analytics approach to building a clinical decision support system for diabetic retinopathy: developing and deploying a model ensemble," *Decision Support Systems,* vol. 101, 2017, pp. 12-27.

[6] X. Ma, Z. Tao, Y. Wang, H. Yu, and Y. Wang, "Long short-term memory neural network for traffic speed prediction using remote micro wave sensor data," *Transportation Research Part C: Emerging Technologies,* vol. 54, 2015, pp. 187 - 197.

[7] H. Jeong, Y Jang, P. Bowman, and N. Masoud, "Classification of motor vehicle crash injury severity: A hybrid approach for imbalanced data," *Accident Analysis & Prevention,* vol. 120, 2018, pp. 250-261.

[8] C. Lin, D. Wu, H. Liu, X. Xia, and N. Bhattarai , "Factor Identification and Prediction for Teen Driver Crash Severity Using Machine Learning: A Case Study," *Applied Sciences,* vol. 10, no. 5, 2020, pp. 1675.

[9] J. Zhang, Z. Li, Z. Pu, and C. Xu, "Comparing prediction performance for crash injury severity among various machine learning and statistical methods," *IEEE Access,* vol. 6, 2018, pp. 60079-60087.

[10] M. Bin and S. Son, "Analysis of factors influencing traffic accident severity according to gender of bus drivers," *J.of Korean Society of Transportation,* vol. 36, no. 6, 2018, pp. 440-451.

[11] T. Chong and B. Kim, "American Sign Language Recognition System Using Wearable Sensors with Deep Learning Approach," *J.of Korea Institute of Electronic Communication Science,* vol. 15, no. 2, 2020, pp. 291-298.

[12] M. Kim, "Variation for Mental Health of Children of Marginalized Classes through Exercise Therapy using Deep Learning," *J.of Korea Institute of Electronic Communication Science,* vol. 15, no. 4, 2020, pp. 725-732.

[13] G. Ke, Q. Meng, T. Finley, T.Wang, W.Chen, W. Ma, and T. Liu, "Lightgbm: A highly efficient gradient boosting decision tree," *Advances in neural information processing systems,* 2017. 9.

[14] H. Trevor, R. Tibshirani, and J. Friedman. *The elements of statistical learning.* Stanford: Springer, 2009, pp. 337-387.

저자 소개

이현미(Hyun-Mi Lee)

2014년 London School of Economics and Political Science(통계석사)

2020년 아주대학교 교통공학과 박사과정 수료
2017년~현재 아주대학교 교통공학과 연구원
※ 관심분야 : ITS통신시스템, ICT, C-ITS

전교석(Gyo-Seok Jeon)

2011년 아주대학교 대학원 교통공학과 졸업(공학석사)

2016년 아주대학교 대학원 교통공학과 졸업(공학박사)
2017년~현재 아주대학교 TOD기반지속가능도시교통연구센터 연구 조교수
※ 관심분야 : ITS통신시스템, ICT, C-ITS

장정아(Jeong-Ah Jang)

2002년 아주대학교 대학원 교통공학과 졸업(공학석사)

2009년 아주대학교 대학원 교통공학과 졸업(공학박사)
2014년~현재 아주대학교 TOD기반지속가능도시교통연구센터 연구교수
※ 관심분야 : ITS통신시스템, ICT, C-ITS