XGBoost를 활용한 이륜자동차 교통사고 심각도 비교분석

Comparative Analysis of Traffic Accident Severity of Two-Wheeled Vehicles Using XGBoost

권 철 우* · 장 현 호**

* 주저자 : 인천대학교 도시융·복합학과 박사과정
** 교신저자 : 인천대학교 도시과학연구원 책임연구원

Cheol woo Kwon* · Hyun ho Chang**

* Department of Urban Convergence Engineering, Incheon National University
** Urban Science Institute, College of Urban Science, Incheon National University

† Corresponding author : Hyunho Chang, nettrek@hanmail.net

Vol.20 No.4(2021)
August, 2021
pp.1~12

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2021.20.4.1

Received 8 June 2021
Revised 5 July 2021
Accepted 23 July 2021

요 약

최근 코로나 19의 영향으로 이륜자동차 교통사고의 발생은 이전보다 급격히 증가하게 되었고 이륜자동차 사고방지를 위한 다각적인 노력이 필요한 시점이다. 본 연구에서는 XGBoost를 활용하여 최근 10년간 발생한 이륜자동차 교통사고 자료로 사고 심각도에 영향인자를 도출하여 각 영향인자가 주는 영향력을 분석하였다. 전체 변수 중 교통사고 심각도에 영향을 주는 변수는 신호 위반을 하였을 경우가 압도적으로 높았으며, 운전자 연령대가 60대 이상일 경우, 이륜자동차 단독사고일 경우, 중앙선 침범 사고일 경우 순으로 높은 것으로 나타났다. 연구 결과를 바탕으로 이륜자동차의 심각한 교통사고의 방지와 안전관리를 강화하기 위한 합리적인 제도 개편방안을 제시하였다.

핵심어 : 이륜자동차 교통사고, 교통사고 심각도, 머신러닝 앙상블

ABSTRACT

Emergence of the COVID 19 pandemic has resulted in a sharp increase in the number of two-wheeler vehicular traffic accidents, prompting the introduction of numerous efforts for their prevention. This study applied XGBoost to determine the factors that affect severity of two-wheeled vehicular traffic accidents, by examining data collected over the past 10 years and analyzing the influence of each factor. Among the total factors assessed, variables affecting the severity of traffic accidents were overwhelmingly high in cases of signal violations, followed by the age group of drivers (60s or older), factors pertaining only to the car, and cases of centerline infringement. Based on the research results, a reasonable legal reform plan was proposed to prevent serious traffic accidents and strengthen safety management of two-wheeled vehicles. Based on the research results, we propose a reasonable legal reform plan to prevent serious traffic accidents and strengthen safety management of two-wheeled vehicles.

Key words : Two-wheeled traffic accident, Severity of accident, XGBoost, SHAP

Ⅰ. 서 론

국내 교통사고 사망자는 2020년 기준 3,079명으로 집계되어 통계 집계를 시작한 이래 최저치를 기록하였다. 경찰청 통계에 따르면 2019년 3,349명에서 2020년 270명이 줄은 3,079명으로 1977년부터 1년 단위로 통계작성을 시작한 이래 가장 낮은 수치다(Korea National Police Agency, 2021). 어린이 보호구역 내 무인단속장비가 확충되고 시내 주요 도로는 50km/h, 동네 도로는 30km/h로 차량 제한속도를 낮추는 '안전운전 5030' 정책이 교통사고로 인한 사망자 감소로 이어진 것으로 분석된다.

하지만 인구 10만 명당 사망자는 OECD 평균(5.6명)보다 높은 5.9명으로 다른 선진국과 비교할 때 여전히 교통안전 수준이 미흡한 수준이다. 특히, 이륜자동차 교통사고가 급격히 증가하고 있는 것으로 집계되었다(MOLIT, 2021). 2020년에 발생한 교통사고 전체 사망자 수는 1년 전보다 줄어들었으나, 이륜자동차 교통사고 사망자 수는 전년 대비 5.4% 증가하여 525명으로 집계되었다. 이륜자동차는 구매 및 유지비용 측면에서 일반 자동차에 비하여 저렴하고 이동 편의성 또한 뛰어남에 따라 소규모 화물 운송에 강점이 있어, 최근 코로나19의 영향으로 퀵서비스나 배달 대행 사업영역이 확대되었고 이륜자동차는 우리의 일상생활과 더욱 밀접한 교통수단이 되고 있다(KOTI, 2020). 2021년 4월 기준으로 등록된 이륜자동차는 약 230만 대로 전체 자동차 중에서 약 9.4%의 비중을 차지하고 있다(MOLIT, 2021). 이렇듯 이륜자동차는 일반 자동차보다 비교적 저렴하고 접근하기 편해 수요가 증가하게 되었고 이륜자동차로 인한 교통사고 발생 또한 증가하게 된 것으로 분석된다.

특히, 이륜자동차의 특성상 일반 자동차와 비교해 운전자를 보호할 차체가 없고 가속도까지 빠르므로 치사율이 높아 문제가 더욱 심각하다. 2019년 기준으로 2.38%인 이륜자동차 사고의 치사율은 전체 교통사고의 치사율인 1.46%보다 약 1.6배 높게 나타나고 있어 이륜자동차 사고방지를 위한 노력이 필요한 상황이다(KOTI, 2020).

이에 발맞춰 정부는 이륜자동차의 불법 주행을 적발하는 '이륜자동차 공익 제보단' 운영과 버스, 택시 등 사업자용 자동차를 활용해 단속을 강화하고 제보를 활성화하는 등 다양한 정책을 제시하여 이륜자동차 교통사고 감소를 위한 적극적인 노력을 기울이고 있다. 단기적으로는 이러한 안전관리 대책이 효과를 거둘 수 있겠지만 과거의 유사 사례에서도 볼 수 있듯이 장기적인 측면에서의 대책으로는 부족한 면이 있으므로, 이륜자동차 교통사고를 줄이기 위한 다각적인 접근과 가시적인 성과를 도출하기 위한 연구가 더욱 필요한 실정이다(KOTI, 2020).

이러한 측면에서 본 연구는 이륜자동차 교통사고 심각도에 영향을 주는 요인을 도출하여 각 요인이 미치는 영향력을 분석하고 이를 활용하여 이륜자동차로 인해 발생하는 교통사고를 예방하기 위한 법규 개편방안을 제시하는 데 그 목적이 있다.

이를 위해 최근 10년간(2010~2019) 전국 단위에서 발생한 이륜자동차 가해 교통사고 자료를 활용하여 분석하였으며, 분석 기법으로는 최근 개발된 머신러닝 기법 중 효율성과 유연성, 휴대성이 뛰어나도록 최적화되어 다양한 현업에서 활용되고 있는 XGBoost를 활용하였다. XGBoost는 Decision Tree 기반 앙상블 머신러닝 알고리즘으로서 Gradient Boosting 프레임워크를 사용하여 정형화된 데이터를 예측할 때 훌륭한 성능을 보이는 알고리즘이다. 회귀모형과 분류모형을 개발하여 우수한 성능을 보일 수 있고, 변수의 중요도를 정량적으로 측정할 수 있으므로 효율적인 분석이 가능하여 본 연구에 활용하였다.

본 연구의 구성은 다음과 같다. 제 2장에서는 관련 이론과 연구 고찰을 설명한다. 제 3장에서는 이륜자동차 교통사고 심각도 분석 방안에 관해 설명하고, 제 4장에서는 분석 결과에 대해 해석한다. 마지막으로 5장에서는 본 연구의 결론과 정책적 제언을 제시한다.

Ⅱ. 관련 이론 및 연구 고찰

1. 머신러닝 알고리즘 - XGBoost

머신러닝을 활용한 분석 방법들은 예측력을 높이기 위해 단일 모델을 사용하기보다 복수의 모델을 활용한 앙상블 학습을 활용할 수 있다. 우수한 성능으로 다양한 분야에서 활용되고 있는 XGBoost 또한 앙상블 학습 중 하나이다. XGBoost는 의사결정나무 모델에 단순한 분류가 가능한 예측 모델들을 결합하여 더욱 강한 예측 모델을 만드는 부스팅 기법을 적용하였다. 주어진 데이터를 분류기를 통해서 학습하고 학습된 결과에서 나타나는 오차를 또 다른 분류기에서 학습하여 오차를 줄인다. 정형 데이터를 예측할 때 훌륭한 성능을 보이는 알고리즘으로서, 회귀 모형 또는 분류 모형에서 활용 가능하다. 기존 모형인 GBM(gradient boosting machine)보다 빠르며, 과적합 방지를 위해 다양한 변수를 조절할 수 있다. 일반적으로 분류와 회귀영역에서 뛰어난 예측성능을 보인다. 2015년 한 해 동안 Kaggle에서 우승한 29개의 과제에서 17개의 모형이 XGBoost를 활용한 것으로 나타났다(Chen and Guestrin, 2016). XGBoost가 다양한 분야의 문제를 해결하는 데 유용하다는 것은 모형의 활용성을 반증할 수 있다. XGBoost는 개발한 모형의 해석을 위해 입력 변수의 중요도를 Shapley value를 통해 수치화하여 해석할 수 있다. Shapley value는 Game Theory의 알고리즘 중 하나로, Game에서 각각의 Player의 기여하는 부분을 계산하는 기법을 말한다. 아래 수식은 Shapley value를 계산하는 수식과 Shapley value를 활용하여 변수의 중요도를 계산하는 수식을 나타낸다.

$$\Phi_j(val) = \sum_{S \subseteq x1,\ldots,x_p/x_j} \frac{|S|!(p-|S|-1)!}{p!}(val(S \cup x_j) - val(S)) \quad \cdots\cdots (1)$$

$$val_x(S) = \int \hat{f}(x_1,\ldots,x_p)dP_{X_2X_4} - E_X(\hat{f}(X)) \quad \cdots\cdots (2)$$

$$I_j = \sum_{i=1}^{n} \left|\Phi_j^{(i)}\right| \quad \cdots\cdots (3)$$

Lee and Sun(2020)은 고속도로의 전략적인 유지관리 계획 수립을 위해 XGBoost를 활용하여 고속도로 포상 파손 예측 모형을 제안하였다. 구축한 데이터 셋의 불균형 문제를 해결하기 위해 언더 샘플링과 오버 샘플링, 그리고 혼합 샘플링 방법을 활용하여 비교하였다. 성능 평가 결과, 오버 샘플링이 포장 파손 예측 성능에 가장 우수한 성능을 보였다. 연구를 통해 장래 고속도로 포장 유지보수 예산의 추정에 중요한 기초정보가 활용될 것으로 기대하였다.

Choi et al.(2020)은 열차의 차상 가속도 데이터를 기반으로 궤도의 품질을 결정하는 지표 중에 하나인 궤도 품질지수를 머신러닝을 활용하여 개발하였다. 서포트 벡터머신과 랜덤포레스트, XGBoost 모형을 개발하여 성능을 비교 평가하였으며 XGBoost가 85% 이상의 가장 높은 예측 정확도를 보인 것으로 분석하였다. 차량 진동 가속도를 이용한 궤도품질지수를 예측하기 위해서는 앙상블 알고리즘을 가지는 모델을 적용하는 것이 적절할 것으로 판단하였다.

Han et al.(2020)은 소비자들이 재무 스트레스를 경험하고 있는지 탐색하여 살펴보고 그들의 영향력을 XGBoost를 활용하여 파악하였다. 온라인 설문조사를 통해 수집된 2,006개의 데이터를 활용하였으며, 분석 결과, 소비자의 주관적 인식이 재무스트레스 수준에 미치는 영향력이 상당히 큰 것으로 나타났다. 특히 단기적인 계획에 바탕을 둔 비상자금과 관련한 요인들의 영향력이 큰 것으로 분석하였다. 연구 결과를 통해 소비자가 경제적 복지를 증진시킬 수 있는 방향을 제시하였다.

2. 교통사고 심각도 추정 기존 연구

교통사고 심각도에 미치는 요인에 관련하여 다양한 연구가 수행되었다.

Park and Shin(2019)의 연구에서는 고속도로 사고 자료와 기상자료를 매칭하여 위계적 순서형 모형을 사용하여 교통사고 심각도에 영향을 미치는 변수들을 분석하였다. 분석 결과 톨게이트 및 램프 구간, 내리막 경사 3% 이상, 콘크리트 방호벽 등이 기상 상태에 따라 사고 심각도에 미치는 영향이 달라지는 것을 확인하였고, 도로기하구조와 기상상태의 복합적인 영향은 강우량 또는 강설량이 선형적이지 않을 수 있음을 확인하였고 분석 결과를 기반으로 안전개선 대책을 제시하였다.

Yoon and Lee(2019)는 2015~2017년의 서울시의 보행자 교통사고를 대상으로 로지스틱 회귀분석을 시행하여 주·야간의 보행자 교통사고 심각도에 영향을 미치는 요인을 살펴보았다. 분석결과, 청소년 및 고령 운전자는 보행자 사고 심각성에 관련이 높은 것으로 나타났고, 교통밀도 또한 사고 심각성에 영향을 주는 것으로 나타났다. 분석 결과를 통해 보행자 교통사고의 심각성을 줄일 수 있는 정책적 시사점을 도출하였다.

Han et al.(2020)은 2017년과 2018년 전국에서 발생한 707건의 PM 사고를 대상으로 순서형 프로빗 모형을 활용하여 사고 심각도를 구분하여 분석하였다. 분석 결과, 도로 및 환경적인 요인으로는 5월, 시간대는 14시와 21시, 23시, 도로가 젖은 상태일 경우, 사고 발생 장소가 교차로인 경우, 70대 운전자인 경우, PM대 차량 사고인 경우 사고 심각도가 높은 것으로 분석하였다. 이 연구가 PM 교통사고 심각도 분석을 위한 기초 자료로 활용될 수 있을 것으로 판단하였다.

Na and Park(2012)는 청주시의 주간선 도로 12개에서 발생한 이륜자동차 교통사고를 순서형 로짓모형을 활용하여 분석하였다. 분석 결과, 여름철에 교통사고가 빈번히 발생하는 것에 반해 겨울철에 사고의 심각도가 높은 것으로 나타났다. 연령의 경우 25세 이하의 연령대에서 사고 심각도가 높았으며, 과속으로 인한 위반에서 사고의 심각도가 높은 것으로 나타났다.

Kim and Park(2019)는 인천광역시에서 발생한 2014~2016년 이륜자동차 교통사고를 토대로 순서형 프로빗 모형을 적용하여 이륜자동차 사고 심각도에 영향을 미치는 요인을 도출하고 분석하였다. 분석결과, 차량 단독사고 중 공작물 충돌과 전도·전복사고, 남성 가해 운전자, 차대 보행자, 차대차의 사고에서 사고 심각도가 높게 나타나 사고 피해를 최소화 할 수 있는 제도적 안전장치와 대책 마련이 필요할 것으로 분석하였다.

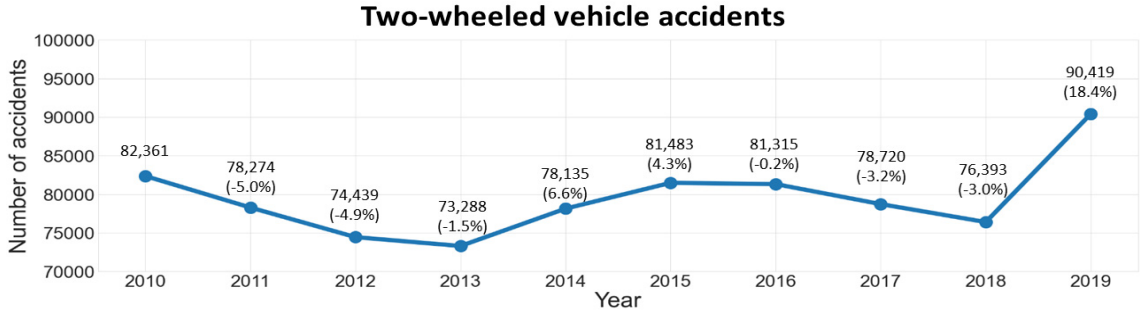
3. 기존 연구와의 차별성

선행 연구를 살펴본 결과, 교통사고 심각도와 같이 순서형 종속 변수를 활용한 모형 개발에서는 순서형 로짓 모형이나 순서형 프로빗 모형을 활용한 연구가 일반적이었다. 이러한 전통적인 통계 모형은 정해진 분포나 가정을 통해 실패 확률을 줄이는 데에 가장 큰 목적이 있다. 따라서 모형의 복잡성보다는 단순성을 추구하기 때문에 모집단의 정규분포, 선형성, 등분산성 등 현실에 적용하기 부적합할 수 있는 무리한 가정이 필요할 수 있다. 이와 달리, 머신러닝은 예측의 성공 확률을 높이는 데에 가장 큰 목적을 두어, 모델의 신뢰도나 정교한 가정은 상대적으로 중요성이 낮아지며, 광활한 데이터를 예측을 수행할 수 있는 장점이 있다. 따라서 본 연구에서는 기존 통계 모형보다는 시스템 성능을 높이기 위해 다양한 분야에서 활용되고 있는 머신러닝을 활용하여 모형을 개발했다는 점에서 기존 연구와의 차별성을 가진다. 특히, 본 연구에서 활용한 XGBoost 모형은 정형 데이터에서 뛰어난 예측성능을 가졌으며, 개발한 boosted tree에서 각 변수가 모형에 얼마나 큰 영향을 미치는지 변수 중요도 함수를 통해 분석할 수 있다는 장점을 가지고 있다.

Ⅲ. 이륜자동차 교통사고 심각도 분석 방안

1. 데이터 수집 및 전처리

본 연구에 사용된 이륜자동차 교통사고 자료는 한국교통안전공단 TAAS에서 제공하는 2010~2019년 전국에서 발생한 10년간의 자료를 수집하여 활용하였다. 수집한 전체 자료 수는 180,862건으로 내용 미기입 등 결측값을 제외하고 전체 자료 중 178,273건을 모형 분석에 활용하였다. 자료를 그래프로 나타내면 <Fig. 1>과 같다. 이륜자동차의 급격한 수요 증가로 2019년 이륜자동차 교통사고는 90,419건으로 전년 대비 18.4%로 크게 증가한 것을 나타났다. 수집한 자료의 특성은 발생일시, 발생지역, 사고내용, 사고유형, 법규위반, 노면상태, 기상 상태, 도로 형태, 운전자 인적사항이 있다. 사고일시는 계절(봄, 여름, 가을, 겨울), 시간(주간, 야간), 주일(평일, 주말)로 분할하였다. 사고의 심각성을 나타내기 위해서는 각 사고의 피해 정도를 하나의 피해 단위로 환산해야한다. 이를 위해서 본 연구에서는 EPDO(대물피해환산계수)[1] 개념을 통해 사고 심각도를 나타냈다. 사고 심각도를 제외한 변수는 범주형 자료이기 때문에 수치형 자료로 변형하기 위해 One hot encoding[2] 방식으로 처리하였으며, 구분한 변수는 아래 <Table 1>와 같다.



<Fig. 1> Number of two-wheeled vehicle accidents

<Table 1> Input data for analysis

Variables		Variable properties
Accident severity(Dependent variable)		EPDO
Personal characteristic variables	Driver age range	0-19, 20-39, 40-59, over 60
	Driver gender	Male, Female
Accident characteristic variables	Accident type	Car to car, Car to pedestrian, Car only, Railroad crossing
	Road type	Intersection, Midblock, Overpass, Underpass, Tunnel, Bridge
	Violation type	Failure to drive safely, Safety distance not secured, Violation of cross traffic rules, Signal violation, Central line infringement, Violation of pedestrian protection obligation, Obstruction of going straight and right turn, Illegal U-turn, Road violation, Speeding too fast
Environmental variables	Season	Spring, Summer, Fall, Winter
	Time of day	Daytime, Nighttime
	Day of week	Weekday, Weekend
	Weather condition	Sunny, Cloudy, Rainy, Fog, Snow
	Road surface condition	Dry, Humid, Frost, Snow, Flooding

1) EPDO(equivalent property damage only) : 사망자수×12 + 중상자수×5 + 경상자수×3 + 부상자수
2) One hot encoding : 자연어 처리를 위해서 문자를 숫자로 처리하는 기법 중 하나

2. 데이터 기초 통계 분석

모형 개발에 앞서 수집한 자료를 통해 기초 통계 분석을 수행하였다. <Table 2>을 살펴보면, 연령대는 60세 이상일 때 35,144명으로 가장 적었지만 사고 심각도가 평균적으로 4.78로 가장 높게 나타났다. 성별은 남성은 169,511명으로 전체 교통사고 중 95.1%로 대부분을 차지하였다. 사고 유형 중 철도 건널목의 사고 심각도가 평균 10.00로 가장 높았다. 도로 유형은 터널이나 교량에서 발생하였을 때 평균 사고 심각도 5가 넘어 높게 나타났다. 법규위반 유형 중 속도위반은 189건으로 가장 적었으나, 사고 심각도가 8.07로 높게 측정되었다. 계절별 이륜자동차 교통사고는 봄은 45,662건, 여름은 50,906건, 가을은 49,897건, 겨울은 31,808건으로 나타났다. 주간보다 야간에 발생한 교통사고가 많았고, 일 대비 교통사고는 평일 25,367건, 주말 25,718건으로 주말이 약간 많았다. 기상 조건은 맑은 날이 158,247건으로 89.3%를 차지하였고, 안개가 꼈을 때 사고 심각도가 평균 5.36으로 가장 높았다. 노면 상태는 건조할 때가 160,577건으로 전체 중 91.0%를 차지하였다.

<Table 2> Descriptive statistical analysis

Variables		Count	Min	Max	Mean	Standard Deviation
Driver age range	0-19	41,259	1	104	4.58	3.12
	20-39	59,752	1	53	4.31	2.88
	40-59	42,118	1	38	4.27	2.72
	over 60	35,144	1	51	4.78	3.19
Driver gender	Male	169,511	1	104	4.47	2.98
	Female	8,707	1	36	4.30	2.71
Accident type	Car to car	125,472	1	104	4.39	3.06
	Car to pedestrian	33,748	1	53	4.39	2.07
	Car only	19,046	1	29	5.05	3.58
	Railroad crossing	7	5	12	10.00	3.42
Road type	Intersection	84,079	1	104	4.48	3.05
	Midblock	6,940	1	46	4.44	2.26
	Overpass	98	1	16	4.59	3.23
	Underpass	451	1	18	4.58	3.05
	Tunnel	307	1	18	5.21	3.45
	Bridge	931	1	32	5.19	3.86
Violation type	Failure to drive safely	95,964	1	53	4.34	2.84
	Safety distance not secured	12,412	1	43	3.98	2.67
	Violation of cross traffic rules	10,973	1	51	4.25	2.76
	Signal violation	30,913	1	104	4.98	3.32
	Central line infringement	9,944	1	94	5.13	3.85
	Violation of pedestrian protection obligation	4,281	1	25	4.43	1.93
	Obstruction of going straight and right turn	5,951	1	24	4.28	2.65
	Illegal U-turn	888	1	46	4.84	3.64
	Road violation	2,121	1	25	4.35	3.01
	Speeding too fast	189	1	36	8.07	5.30
Season	Spring	45,662	1	53	4.50	2.99
	Summer	50,906	1	94	4.48	3.02
	Fall	49,897	1	41	4.50	2.96
	Winter	31,808	1	104	4.30	2.88
Time of day	Daytime	82,163	1	53	4.40	2.94
	Nighttime	96,110	1	104	4.51	3.00
Day of week	Weekday	126,836	1	104	4.40	2.88
	Weekend	51,437	1	94	4.60	3.19
Weather condition	Sunny	158,247	1	104	4.46	2.97
	Cloudy	7,337	1	94	4.66	3.38
	Rainy	11,116	1	32	4.28	2.72
	Fog	144	1	22	5.36	3.52
	Snow	412	1	12	4.06	2.33
Road surface condition	Dry	160,577	1	104	4.48	2.99
	Humid	15,147	1	32	4.31	2.80
	Frost	554	1	34	4.04	2.81
	Snow	172	1	12	3.85	2.30
	Flooding	4	3	5	3.50	1.00

3. 구축 데이터 샘플

구축한 자료의 예시는 아래 <Fig. 2>과 같다. 교통사고는 시간과 장소가 불특정하게 발생하기 때문에 교통사고 발생 시 기상상태, 도로유형, 법규위반 등 대표되는 유형이 없어 특성을 '기타'로 분류하는 경우가 존재한다. 본 연구에서는 사고 심각도의 영향을 주는 요인을 파악하는 연구 목적을 위해 '기타'로 분류된 열은 활용하지 않고 총 178,273행×45열의 데이터 셋을 구축하여 연구에 활용하였다.

	Accident Severity	Season_Fall	Season_Spring	Season_Summer	Season_Winter	Time of day_Daytime	Time of day_Nighttime	Day of week_Weekday	Day of week_Weekend	...
0	1	1	0	0	0	1	0	1	0	
1	3	0	0	1	0	0	1	1	0	...
2	3	0	1	0	0	1	0	0	1	
3	6	1	0	0	0	1	0	1	0	
4	6	0	1	0	0	1	0	1	0	...
...	...	...	...	...	...	...	...	...	...	
178268	5	1	0	0	0	1	0	1	0	
178269	5	0	0	1	0	1	0	0	1	...
178270	3	0	0	1	0	0	1	0	1	
178271	12	0	1	0	0	1	0	0	1	
178272	5	1	0	0	0	1	0	1	0	...

178273 rows × 45 columns

<Fig. 2> Processed data sample

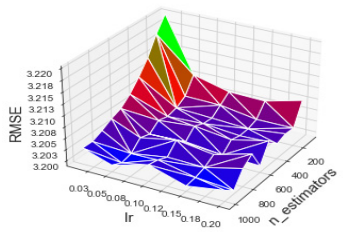
4. 모형 성능 평가 지표 선정

본 연구에서 개발한 모형의 성능을 평가하기 위해서는 성능평가 지표 선정이 필요하다. 일반적으로 회귀모형에서 사용되는 성능 평가 지표로는 MAE, MSE, RMSE, RMSLE 등이 있다. 본 연구에서는 큰 오류 값 차이에 대해서 크게 패널티를 부여하고, 오류 지표를 실제 값과 유사한 단위로 다시 변환하기 때문에 해석하기 쉬운 RMSE를 활용하여 성능을 평가하였다. RMSE의 수식은 아래와 같다.

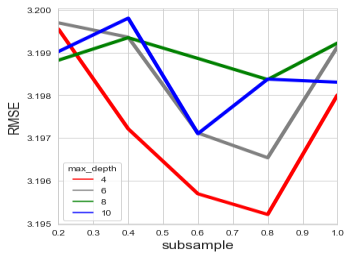
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y_i})^2} \quad \cdots\cdots (4)$$

5. 모형 최적화

본 연구의 분석은 파이썬 3.7를 수행되었으며, 178,273건의 자료 중 70%인 124,791건으로 학습하고, 나머지 30%인 53,482건으로 평가를 수행하였다. XGBoost의 파라미터 최적화 과정은 크게 두 번으로 나누어 lr(learning rate)과 n_estimators의 최적화, max_depth과 subsample의 최적화 과정으로 진행하였다. RMSE를 통해 평가하였으며, 모든 최적화 프로세스의 안정적인 결과를 위해 5-fold cross-validation을 거쳤다. 첫 번째 과정을 위해 lr은 {0.01, 0.05, 0.1, 0.2}로, n_estimators는 {50, 100, 150, ..., 1000}로 설정하였고, Grid search 분석 결과, {lr, n_estimators}={0.05, 900}에서 가장 RMSE가 낮은 결과를 보였다. 두 번째 과정에서 max_depth는 {4, 6, 8, 10}으로 subsample은 {0.4, 0.6, 0.8, 1.0}으로 설정하였고 Grid search 분석 결과, {max_depth, subsample}={4, 0.8}에서 RMSE 최적화 값을 가진 것으로 확인되었다.



<Fig. 3> Optimization of lr and n_estimators



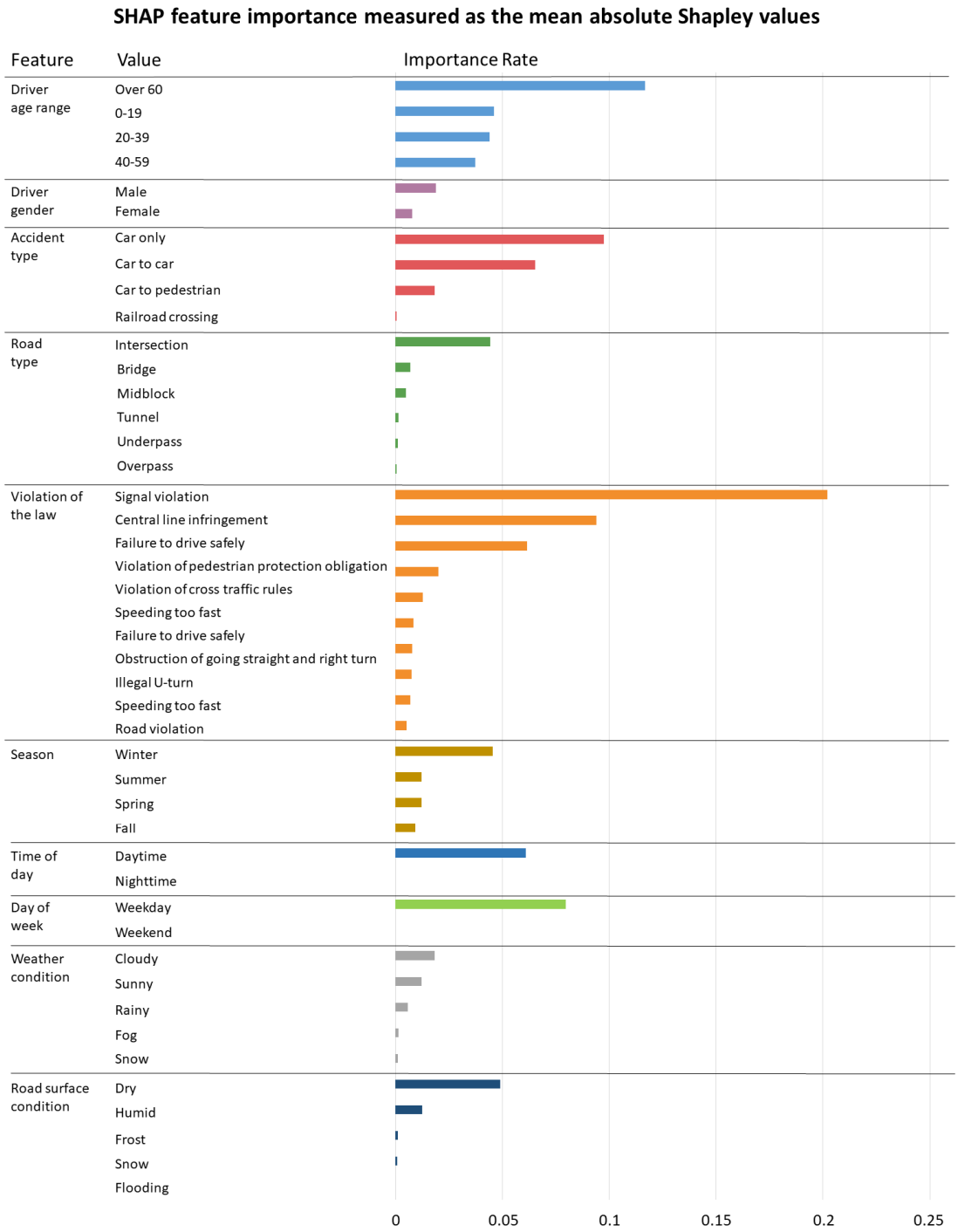
<Fig. 4> Optimization of max_depth and subsample

Ⅳ. 분석 결과

최종적으로 XGBoost을 활용하여 이륜자동차 교통사고 심각도를 분석한 결과, RMSE 값이 2.951로 측정되었다. 성능 평가를 위해 Decision tree와 Random forest를 개발한 모형의 성능은 RMSE 값이 각각 3.120, 3.105로 XGBoost를 활용한 모형의 성능이 가장 우수한 것으로 나타났다. 개발한 XGBoost를 활용한 사고 심각도 모형은 SHAP 패키지를 활용하여 영향을 미치는 요인을 수치화하여 분석하였다. 교통사고 심각도에 가장 큰 영향을 주는 변수는 전체 변수 중 신호 위반을 하였을 경우가 변수 중요도가 0.202로 측정되어 가장 높았으며, 운전자 연령대가 60대 이상일 경우, 이륜자동차 단독사고일 경우, 중앙선 침범 사고일 경우 순으로 높은 것으로 나타났다. 운전자 연령대 변수 중에서는 60대 이상일 경우는 다른 연령대보다 변수 중요도가 높아 교통사고 심각도에 큰 영향을 주는 것으로 나타났다. 운전자 성별은 남성이 0.019로 여성(0.008)보다 2배 정도 높았다. 사고 유형에서 차량 단독으로 사고가 발생하였을 경우가 0.097로 변수 중요도가 가장 높았고, 차대차 사고, 차대사람 사고, 철도건널목 사고 순으로 나타났다. 도로 유형을 살펴보면, 교차로에서 교통사고가 발생하였을 경우 사고 심각도가 0.044로 도로 유형 중 가장 큰 영향을 주는 것 분석됐다. 법규위반 변수 중 신호 위반이 0.202로 전체 변수 중 가장 큰 영향을 주는 것으로 분석되었다. 다음으로 중앙선 침범, 안전운전 불이행, 보행자 보호법 위반 순으로 변수 중요도가 높게 측정되었다. 계절을 살펴보면 겨울(0.045)이 다른 계절보다 교통사고 심각도가 높게 분석됐다. 주간 발생(0.061)이 야간 발생(0.000)보다 교통사고의 심각도에 영향을 주는 것으로 나타났고, 평일(0.080)이 주말(0.000)보다 높은 것으로 나타났다. 기상 상태는 흐린 날씨의 심각도가 0.018로 높았으나, 다른 변수보다 교통사고 심각도의 영향이 낮았다. 노면 상태는 보통 상태인 건조 상태(0.049)일 때가 가장 높게 측정됐고, 젖은 상태(0.012)가 다음으로 높았다.

<Table 3> Feature importance

Variables		Feature importance	Variables		Feature importance
Driver age range	0-19	0.046	Season	Spring	0.012
	20-39	0.044		Summer	0.012
	40-59	0.037		Fall	0.009
	over 60	0.117		Winter	0.045
Driver gender	Male	0.019	Time of day	Daytime	0.061
	Female	0.008		Nighttime	0.000
Accident type	Car to car	0.065	Day of week	Weekday	0.080
	Car to pedestrian	0.018		Weekend	0.000
	Car only	0.097	Weather condition	Sunny	0.012
	Railroad crossing	0.000		Cloudy	0.018
Road type	Intersection	0.044		Rainy	0.006
	Midblock	0.005		Fog	0.001
	Overpass	0.000		Snow	0.001
	Underpass	0.001	Road surface condition	Dry	0.049
	Tunnel	0.001		Humid	0.012
	Bridge	0.007		Frost	0.001
Violation type	Failure to drive safely	0.061		Snow	0.001
	Safety distance not secured	0.007		Flooding	0.000
	Violation of cross traffic rules	0.013			
	Signal violation	0.202			
	Central line infringement	0.094			
	Violation of pedestrian protection obligation	0.020			
	Obstruction of going straight and right turn	0.008			
	Illegal U-turn	0.007			
	Road violation	0.005			
	Speeding too fast	0.008			



<Fig. 5> Feature importance

Ⅴ. 결 론

최근 코로나 19의 영향으로 퀵서비스나 배달 대행 사업영역이 확대되었고, 이륜자동차는 우리의 일상생활과 더욱 밀접한 교통수단이 되었다. 그 영향으로 이륜자동차 사고의 발생은 이전보다 급격히 증가하게 되었고 이륜자동차 사고방지를 위한 다각적인 노력이 필요한 시점이다. 이에 따라 본 연구에서는 XGBoost를 활용하여 2010년부터 2019년까지 발생한 이륜자동차 교통사고 자료로 사고 심각도에 영향을 주는 요인을 도출하여 각 요인이 미치는 영향력을 분석하였다. 수집한 이륜자동차 교통사고 자료 총 178,273건 중 70%를 학습 데이터로 30%를 평가 데이터로 활용하였으며, lr, n_estimators, maxdepth, subsample을 통해 최적화 과정을 거쳤다. 모든 최적화 프로세스는 안정적인 결과를 위해 5-fold cross-validation을 활용하였다. 최종 파라미터는 lr은 0.05, n_estimators은 900, maxdepth은 4, subsample은 0.8로 설정하였고 RMSE는 2.951로 측정되었다. 최종 구축한 XGBoost 모형은 SHAP 패키지를 활용하여 이륜자동차 교통사고 심각도에 영향을 주는 변수의 중요도를 수치화하여 해석하였다. 전체 변수 중 신호 위반을 하였을 경우가 변수 중요도가 0.202로 측정되어 가장 높았으며, 운전자 연령대가 60대 이상일 경우(0.117), 이륜자동차 단독사고일 경우(0.097), 중앙선 침범 사고일 경우(0.094) 순으로 높은 것으로 나타났다. 다양한 원인으로 발생하는 이륜자동차의 심각한 교통사고를 방지하기 위해서는 철저한 안전관리를 위한 합리적인 해결방안의 제시가 필요할 것이다. 본 연구에서 살펴보았던 이륜자동차 교통사고 심각도에 영향을 미치는 요인을 통한 정책적 시사점을 도출하고자 한다.

먼저, 신호위반, 중앙선 침범은 '인명의 보호'를 위협하는 행위로 매우 강력한 제재가 이루어지고 있다. 하지만 이륜자동차의 운전자는 일반 자동차보다 쉽게 적색 신호를 위반하거나 중앙선을 침범하게 되고 이로 인해 발생한 교통사고의 심각성은 높아지는 것으로 판단된다. 이륜자동차용 단속 시스템의 추가적인 운영과 이륜자동차 사고 심각도를 줄일 수 있는 보행자 안전펜스 설치 확대를 통해서 이륜자동차의 법규위반을 줄이기 위한 노력이 필요하다. 두 번째, 고령 운전자는 안전 불감증과 위험 상황에 대한 대처 능력이 떨어지게 되고 이로 인해 사고 심각도가 높다. 고령 운전자 사망사고 중 이륜자동차 교통사고가 3건 중 1건이었으며 이 중, 고령 운전자의 절반 정도(53.6%)만이 안전모를 착용한 것으로 나타났다(KOROAD, 2016). 고령 운전자를 위한 이륜자동차 안전 교육 프로그램의 개발과 보급이 필요하다. 합리적인 이륜자동차 안전 교육 프로그램의 개발과 교육의 효과적인 보급이 이루어지기 위해서 이륜자동차 안전 교육의 관리 담당 정부 기관을 정하여 관련 절차를 법제화하는 것이 매우 중요할 것으로 판단된다. 세 번째, 이륜자동차는 낮은 연령의 접근성이 좋으므로 운전 미숙으로 인한 차량 단독 사고로 이어지는 경우가 많다. 한국교통연구원 연구에 따르면, 차량 단독 사고의 무면허 운전자의 빈도는 다른 사고에 비해 1.5배 높게 나타났다(KOTI, 2014). 이러한 이륜자동차 단독사고 감소를 위한 첨단기술의 적용을 위한 정부의 적극적 지원이 필요할 것이다. 오토바이의 속도를 자동으로 조정하여 앞 차량과 안전한 거리를 유지해주는 ARAS(Advanced rider assistance system)나 하이테크 헬멧 등 이륜자동차의 최신 기술이 적용되어 나아가 궁극적으로 교통안전이 증진될 수 있도록 정부 차원에서의 적극적 지원이 필요한 시점이다.

ACKNOWLEDGEMENTS

본 연구는 국토교통부 교통물류연구사업의 연구비지원(21TLRP-B148966-04)에 의해 수행되었습니다.

REFERENCES

Annie R. R., Srihari P. and Meena P.(2020), "Prediction of Road Accident Severity Using Machine Learning Algorithm," *International Journal of Advanced Science and Technology*, vol. 29, no. 6, pp.116-120.

Chen T. and Carlos G.(2016), "XGBoost: A scalable tree boosting system," *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.785-794.

Chen T. and Guestrin C.(2016), "XGBoost: A scalable tree boosting system," *In Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining. San Francisco*, CA, USA, pp.785-794.

Choi C. Y., Kim H. K., Kim Y. C. and Kim S. S.(2020), "Prediction of Track Quality Index (TQI) Using Vehicle Acceleration Data based on Machine Learning," *Journal of Korean Geosynthetics Society*, vol. 19, no. 1, pp.45-53.

Han D. J., Kim E. C. and Ji M. K.(2020), "Analysis of Severity Factors in Personal Mobility (PM) Traffic Accidents," *Journal of Korean Society of Transportation*, vol. 38, no. 3, pp.232-247.

Jessica P. N., Montserrat G. and Manuela A.(2019), "Predicting Motor Insurance Claims Using Telematics Data-XGBoost versus Logistic Regression," *MDPI*, vol. 7, no. 70, pp.1-16.

Jun M., Yuexiong D., Jack C. P., Yi T., Vincent J. L. and JingCheng Z.(2019), "Analyzing the Leading Causes of Traffic Fatalities Using XGBoost and Grid-Based Analysis: A City Management Perspective," *IEEE*, vol. 7, pp.148062-148072.

Kim T. H. and Park M. H.(2019), "Risk Factors Related to Accident Severity of Two-wheeled Vehicles Using Ordered Probit Model: A Case of Incheon Metropolitan City," *International Journal of Highway Engineering*, vol. 21, no. 2, pp.71-79.

Korea National Police Agency(2021), *Traffic accident statistics*.

KOROAD(2016), *Traffic accident characteristics of elderly drivers over the past 5 years*.

KOTI(2014), *Legal System Improvement Plans for Accident Prevention and Safety Management of Motorcycle in Korea*.

KOTI(2020), *KOTI Special Edition 05"*, pp.24-29.

Lee Y. J. and Sun J. W.(2020), "Predicting Highway Concrete Pavement Damage using XGBoost," *Journal of Construction Engineering and Management*, vol. 21, no. 6, pp.46-55.

MOLIT(2021), *Total Registered Motor Vehicles*.

Mussone L., Bassani M. and Masci P.(2017), "Analysis of factors affecting the severity of crashes in urban road intersections," *Accident Anal. Prevention*, vol. 103, pp.112-122.

Na H. and Park B. H.(2012), "Analysis on the Accident Severity of Motorcycle Using Ordered Logit Model," *Korea Planning Association*, vol. 47, no. 4, pp.233-240.

Park J. W. and Shin C. S.(2019), "A Study on Comparison of the Machine Learning Models for the Trip Distance Prediction of the Seoul Public Bike Sharing Service," *Journal of Knowledge Information Technology and Systems*, vol. 14, no. 2, pp.625-634.

Park S. J., Kho S. Y. and Park H. C.(2019), "The Effects of Road Geometry on the Injury Severity of Expressway Traffic Accident Depending on Weather Conditions," *Journal of the Korea*

Institute of Intelligent Transport Systems, vol. 18, no. 2, pp.12-28.

Yoon J. H. and Lee S. G.(2019), "Comparative Analysis of Factors Affecting the Severity of Pedestrian Crash by Daytime and Nighttime in Seoul, Korea," *Journal of Korea Planning Association*, vol. 54, no. 7, pp.70-88.

Zhang D., Qiu R. W., Deng Y., Ji D. and Li T.(2019), "Novel framework for image attribute annotation with gene selection XGBoost algorithm and relative attribute model," *Appl. Soft Comput.*, vol. 80, pp.57-79.