Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2022.45.2.020

Vol. 45, No. 2 : 20-29, June 2022
ISSN : 2005-0461 / EISSN : 2287-7975

Development of Traffic Accident Prediction Model Based on Traffic Node and Link Using XGBoost

Un-Sik Kim · Young-Gyu Kim · Joong-Hoon Ko†

VAIV Company Inc.

XGBoost를 이용한 교통노드 및 교통링크 기반의 교통사고 예측모델 개발

김운식 · 김영규 · 고중훈†

주식회사 바이브컴퍼니

This study intends to present a traffic node-based and link-based accident prediction models using XGBoost which is very excellent in performance among machine learning models, and to develop those models with sustainability and scalability. Also, we intend to present those models which predict the number of annual traffic accidents based on road types, weather conditions, and traffic information using XGBoost. To this end, data sets were constructed by collecting and preprocessing traffic accident information, road information, weather information, and traffic information. The SHAP method was used to identify the variables affecting the number of traffic accidents. The five main variables of the traffic node-based accident prediction model were snow cover, precipitation, the number of entering lanes and connected links, and slow speed. Otherwise, those of the traffic link-based accident prediction model were snow cover, precipitation, the number of lanes, road length, and slow speed. As the evaluation results of those models, the RMSE values of those models were each 0.2035 and 0.2107. In this study, only data from Sejong City were used to our models, but ours can be applied to all regions where traffic nodes and links are constructed. Therefore, our prediction models can be extended to a wider range.

Keywords : Traffic Accident Prediction, XGBoost, Traffic Links, Traffic Node

1. 서 론

대한민국 경찰청 통계에 따르면 자동차 등록대수가 2010년 약 17.9백만 대에서 2019년 23.6백만 대로 연평균 3.13% 증가한 반면에, 자동차 1만 대당 사고건수는 2010년 126.5건에서 2019년 97건으로 연평균 2.84%, 자동차 1만 대당 사망자수는 2010년 3.1명에서 2019년 1.4명으로 연평균 8.16% 오히려 감소하였다[11]. 대한민국정부는 교통사고 예방을 위해 다양한 교통안전 정책을 꾸준히 추진하여 자동차 1만 대당 사고건수와 사망자수는 지속적으로 감소하고 있다. 하지만, 2018년 기준 OECD 회원국들의 자동차 1만 대당 사망자수 평균값 0.8명과 비교하면 약 75% 높은 수치를 보인다. 일반적으로 교통사고는 다양한 요인들의 복합적인 작용에 의하여 발생하기 때문에 정부는 지속적으로 다양한 교통사고 발생요인을 줄이는 노력이 필요하다.

교통사고 발생요인 관련 연구는 거시적 관점과 미시적 관점으로 구분할 수 있다. 선행연구는 주로 미시적 관점의

Received 23 March 2022; Finally Revised 3 May 2022; Accepted 6 May 2022
† Corresponding Author : joonghoon.ko@vaiv.kr

교통사고 발생요인을 도로의 특정 구간 혹은 특정 지점을 대상으로 분석하였다. 선행연구에서 밝힌 선행연구의 한계점은 다음과 같다. Ryu[17]은 딥러닝을 이용한 고속도로 교통사고 예측모델 개발 연구에서 자료구축 단위인 고속도로 콘존은 지방부 경우에는 구간길이가 길고 다양하여 딥러닝 수행에 필요한 데이터수를 충분히 확보하지 못하는 한계가 존재한다고 하였다. 또한, 세부 기하구조, 시계열 교통량 및 속도 자료 등을 입력변수에 활용한다면 한계점을 해결할 수 있을 것이라 밝혔다[17]. Park[16]은 워싱턴주 내 7개 고속도로를 대상으로 확률적 모수를 이용한 음이항 모형 개발에서 독립변수에 사용된 기하구조의 경우 매년 변화를 추적해야만 하는 자료수집의 어려움이 있다고 밝혔다[16].

본 연구에서는 선행연구에서 한계점으로 밝힌 도로의 세부 기하구조, 시계열 교통량 자료를 활용하면서 매년 자료의 변화를 추적할 수 있는 모델을 개발하기 위해 한국의 국가교통정보센터에서 구축한 교통노드 및 교통링크 데이터, 교통소통정보를 사용하여 모델의 지속성과 확장성을 만족하는 교통사고 예측모델 개발을 위한 데이터셋을 구축하고, Gradient Boosting Machine(GBM) 대비 빠른 수행시간, 과대적합 규제(overfitting regularization), 조기 종료(early stopping), 분류 및 회귀영역에서 예측 성능이 우수한 XGBoost(eXtreme Gradient Boosting)를 이용한 교통노드 및 교통링크 기반의 교통사고 예측모델을 제시하고자 한다.

2. 이론적 배경

2.1 선행연구 고찰

Cho[2]는 도로 네트워크를 따른 교통사고 핫스팟의 시각화 연구에서 10m 단위의 일정한 도로링크별 교통사고 건수를 추출한 방법을 통해 주요 도로 구간 내에서 특히 교차로나 다른 도로로 진입하는 구간에 사고가 많이 발생하는 것을 밝혔다.

Im[3]은 확률모수를 이용한 교통사고 건수 예측모형 개발 연구에서 기존의 음이항 모형보다 확률 모형에서 교통사고 건수를 정확하게 예측할 수 있다고 밝혔다.

Jeong[14]은 도시고속도로의 교통사고 영향요인 분석에 관한 연구를 통해 판별모델에서 공통적으로 나타난 영향요인들 중 환경요인들로 흐린 기상상태, 단일로, 터널 내부, 평지의 커브 및 곡각 구간이 교통사고의 주된 요인으로 작용하고 있다고 밝혔다[4].

Kang[5]은 서울권, 수도권, 부산권에 위치한 4지교차로 106개 지점을 대상으로 퍼지추론 및 신경망 이론을 적용하여 교통 사고모형을 개발하였다. 그 결과 부도로의 교통량, 주도로의 차로수, 교차로 넓이, 부도로 딜레마구간의 길이 등이 사고에 영향을 미치는 요인임을 밝혔다.

Kang[6]은 의사결정나무와 시공간 시각화를 통한 서울특별시 교통사고 심각도 요인 분석에서 심각한 교통사고로 이어지는 경우를 살펴보면 차대사람 또는 차량단독 사고는 고속도로나 특별·광역시도와 같이 폭원이 넓고, 차량속도가 높은 곳에서 승합차나 화물차에서 중상의 교통사고가 일어날 가능성이 높다고 밝혔다.

Lee[12]는 도로위의 기상요인이 교통사고에 미치는 영향을 로지스틱 회귀모형과 의사결정나무모형으로 연구한 결과 기상요인 중에서 강수유무와 기온이 교통사고 발생에 영향을 미치는 요인으로 나타났다고 밝혔다.

Lee[13]는 2007년부터 2012년까지 대전광역시 내 89개 교차로를 대상으로 사고예측 모형을 개발하였다. 연평균 일교통량, 제한속도, 차로수, 우회전 전용차로 설치유무 등이 유효한 설명변수임을 밝혔다.

Oh[15]은 도로선형설계요소의 표준편차를 이용한 설계일관성과 교통사고와의 상관성 연구에서 피어슨 상관분석 결과를 보면 곡선반경, 곡선길이, 편경사크기, 관찰속도 크기가 교통사고와 관계가 있다고 밝혔다.

Park[16]는 워싱턴주 7개의 고속도로를 대상으로 확률적 모수를 이용한 음이항모형을 개발하였다. 그 결과 교통량, 차선수, 길어깨 폭, 횡단곡선 개수, 최대종단구배가 교통사고에 영향을 미치는 요인임을 밝혔다.

Ryu[17]은 딥러닝을 이용한 고속도로 교통사고 예측모델 개발 연구에서 음이항 회귀모형, 포아송 회귀모형, 노출계수를 활용하는 모형보다 딥러닝을 활용한 모형이 예측 신뢰도를 더욱 증가 시킨다고 밝혔다.

2.2 XGBoost(eXtreme Gradient Boosting)

Chen and Guestrin[1]은 훈련속도 및 안정성 향상은 물론 과대적합(overfitting)을 해결하기 위해 Gradient Boosting 알고리즘인 XGBoost(eXtreme Gradient Boosting)를 소개하였다. 이 알고리즘은 병렬처리와 하드웨어 최적화로 빠른 속도를 지원하고 회귀와 분류를 모두 지원하는 모델로 의사결정나무로 구성된 다른 트리기반 앙상블 모델과 달리 CART(Classification and Regression Trees) 모델을 기반으로 구성되어 있다. CART는 함수 형태를 가정하지 않은 회귀모형으로 정의되며 아래와 같은 수식으로 표현할 수 있다.

$$Y' = a * tree_A + b * tree_B + c * tree_C + \ldots$$

여기서, Y ′ 는 타겟(Y)에 대한 예측값, a,b,c,...는 각 트리

A,B,C...에서 나온 가중치를 말한다. 위 수식을 XGBoost의 Gradient Boosting Tree에서 사용하면 아래와 같이 표현될 수 있다[1].

$$Y'_i = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F$$
$$obj(\theta) = \sum_{i=1}^{n} l(y_i, y'_i) + \sum_{k=1}^{K} ohm(f_k)$$

where y'_i = *predict score corresponding*
f_k = *k th decision tree* $\in$ *function space* F
l = *loss function*
ohm = *regularization function*

즉, 여러 개의 트리를 사용하여 학습하고, 각 트리의 결과 값의 합을 예측값으로 사용한다. 이러한 방법은 과대적합은 물론 기존 GBM이 가지고 있는 취약점을 보완할 수 있다.

2.3 SHAP(SHapley Additive exPlanations)

최근 AI 모델이 특정 결정을 내린 원인과 그 작동 원리를 사람들이 쉽게 파악하고 이해할 수 있는 형태로 설명 가능한 인공지능(eXplainable AI, XAI)이 대두되었다. Lundberg et al.[14]은 개별적인 의사결정을 설명하기 위해 Shapley Value를 이용하는 방법을 제안하였다. 특성의 부정적인 영향이 반영되지 않는 주요변수 도출방법에 비해 더욱 정확한 영향력을 표시해 준다. Shapley Value는 게임이론(Game Theory)으로부터 출발한 알고리즘으로 특정 변수가 예측력에 얼마나 기여하는지 파악하기 위해 특정 변수와 관련된 모든 변수의 조합들을 입력시켰을 때 나온 결과값과 비교하면서 변수의 기여도를 계산한다. 이에 대한 수식은 아래와 같이 정의할 수 있다.

$$\Phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z' \setminus i)]$$

Φ_i = *shapley value of attributes i*
n = *total number of attributes*
$f_x(z')$ = *contribution of all attributes*
$f_x(z' \setminus i)$ = *all other attributes except attributes i contribution obtained using*

즉, attribute i의 기여도는 전체 기여도 중에서 attribute i를 제외한 기여도를 뺀 값이다. 이와 같은 방법으로 독립변수가 종속변수에 미치는 긍정적인 영향과 부정적인 영향을 모두 반영한 평균 기여도를 도출할 수 있다[14].

3. 자료 구축

3.1 자료 수집

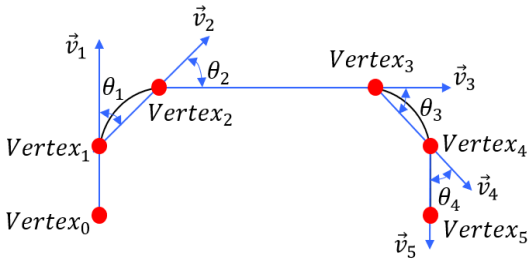
본 연구에서는 교통사고정보, 교통노드정보, 교통링크정보, 교통노드 회전정보, 세종특별자치시 교통소통정보, 대전광역시 기상정보, 수치표고모델(Digital Elevation Model, DEM)을 사용하였다. 또한, 경찰청 교통사고분석시스템(Traffic Accident Analysis System, TAAS)으로부터 2017년부터 2019년까지 세종특별자치시 교통사고정보를 수집하였다[10]. 교통노드정보, 교통링크정보, 교통노드 회전정보는 국가교통정보센터에서 제공하는 2021년 6월 정보를 수집하였다[9]. DEM 정보는 미국지질조사국(United States Geological Survey, USGS)에서 제공하는 30M 규격의 세종특별자치시 DEM 데이터를 수집하였다[18]. 세종특별자치시 교통소통정보는 국가교통정보센터에서 제공하는 2020년 1월부터 2021년 5월까지의 자료를 수집하였다[9]. 대전광역시 기상정보를 위해서는 기상청에서 제공하는 종관기상관측자료(ASOS) 중에서 대전(133번) 자료에서 2017년부터 2019년까지의 데이터를 수집하였다[7].

3.2 신규변수 생성

교통노드 기본정보는 연결된 교통링크 식별자와 교통노드의 꼭지점(Vertex) 정보이며, 교통링크 기본정보는 교통링크 시작지점, 교통링크 종료지점, 차로수, 도로등급, 도로유형, 제한속도, 길이, Vertex 정보이다. 본 연구에서는 교통노드 및 교통링크의 기본정보만으로 교통노드와 교통링크가 갖는 기하구조를 파악할 수 없기 때문에 다른 데이터를 이용하여 교통노드와 교통링크의 기하구조를 표현할 수 있는 신규 변수를 생성하였다.

교통노드는 교통링크 시작지점과 교통링크 종료지점을 이용하여 다수의 교통링크와 연결될 수 있다. 해당 정보를 이용하면 노드에 연결된 링크수, 노드에 연결된 진입 링크수, 노드에 연결된 진출 링크수, 진입 차로수와 진출 차로수의 차이, 진입 링크의 차로수 합, 진출 링크의 차로 수 합, 연결된 링크의 평균 제한속도, 연결된 링크의 제한속도, 변동계수 정보를 생성할 수 있다.

교통링크와 교통노드 데이터에는 Vertex 정보가 포함되어 있다. 교통노드의 Vertex 정보는 교통노드의 위치정보가 되며, 교통링크의 Vertex 정보는 링크를 이루는 각 꼭지점의 위치정보이다. 해당 정보를 이용하면 특정 교통링크의 평균곡률을 구할 수 있다(<Figure 1> 참조).

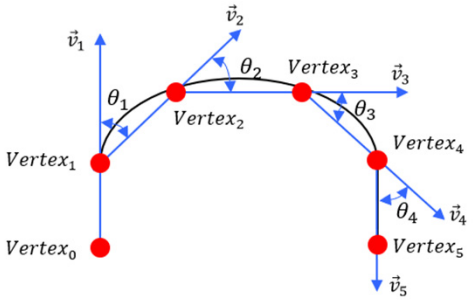


<Figure 1> Traffic Link Configuration Example

<Figure 1>에서 교통링크의 연속되는 3개의 $Vertex_1$, $Vertex_2$, $Vertex_3$ 이 생성하는 $\overrightarrow{v_1}$과 $\overrightarrow{v_2}$가 이루는 각 θ_1을 구할 수 있다. 같은 방법으로 θ_1, θ_2, θ_3, θ_4를 구하면 아래 수식을 이용하여 교통링크의 평균곡률을 구할 수 있다.

$$Average\ curvature = \frac{\sum_{i=1}^{n(Vertex)-2} \theta_i}{n(Vertex)}$$

교통링크는 도로의 곡선부를 각 Vertex의 좌표값으로 표현한다. $Vertex_n$의 곡선부 유형은 θ_{n-1}, θ_n, θ_{n+1} 각각의 값이 존재하는지와 임계값 θ_α보다 크거나 작은지를 판단하여 해당 Vertex에 대해 진입부(Entry Point), 회전부(Rotating Point), 진출부(Exit Point), 직선부 여부를 판단할 수 있다. 임계값 θ_α는 직선부의 θ_i의 값이 곡선구간이라 판단하는 임계값이다. 본 연구에서는 임계값 θ_α=1을 사용



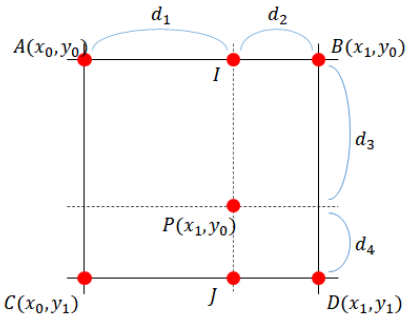
$Vertex_n$	Type	θ_{n-1}	θ_n	θ_{n+1}
$Vertex_0$	Entry point	None or $\theta_{n-1} < \theta_\alpha$	None or $\theta_n < \theta_\alpha$	$\theta_{n-1} > \theta_\alpha$
$Vertex_1$	Entry point	None or $\theta_{n-1} < \theta_\alpha$	$\theta_n > \theta_\alpha$	$\theta_{n-1} > \theta_\alpha$
$Vertex_2$	Rotating point	$\theta_{n-1} > \theta_\alpha$	$\theta_n > \theta_\alpha$	$\theta_{n-1} > \theta_\alpha$
$Vertex_3$	Rotating point	$\theta_{n-1} > \theta_\alpha$	$\theta_n > \theta_\alpha$	$\theta_{n-1} > \theta_\alpha$
$Vertex_4$	Exit point	$\theta_{n-1} > \theta_\alpha$	$\theta_n > \theta_\alpha$	None or $\theta_{n-1} < \theta_\alpha$
$Vertex_5$	Exit point	$\theta_{n-1} > \theta_\alpha$	None or $\theta_n < \theta_\alpha$	None or $\theta_{n-1} < \theta_\alpha$

θ_α = Threshold

<Figure 2> Traffic Link Curve Type

하였다. 진입부, 회전부, 진출부 모두 해당하지 않는다면 직선부이다. 각 Vertex의 진입부, 회전부, 진출부를 구한 뒤 진입부와 진출부 쌍의 개수를 교통링크의 곡선구간 개수로 사용한다(<Figure 2> 참조).

본 연구에서의 고도정보는 30M 간격의 DEM 정보를 사용하였다. 30M 간격의 고도정보는 쌍선형보간법을 이용하여 각 Vertex의 고도값을 보다 정확하게 추정할 수 있다. 쌍선형보간법은 1차원 공간에서 사용하는 선형보간법을 2차원 공간에서 사용하기 위해 확장한 보간법이다. <Figure 3>은 쌍선형보간법을 이용하여 $f(p)$를 도출하는 방법을 나타낸다. 먼저 선형보간법을 통해 I와 J의 좌표값과 고도값을 구한 뒤 I와 J의 좌표값과 고도값을 이용하여 선형보간법을 통해 P의 고도값을 구한다.



$$f(I) = \frac{d_1}{d_1 + d_2} f(B) + \frac{d_1}{d_1 + d_2} f(A)$$

$$f(J) = \frac{d_1}{d_1 + d_2} f(D) + \frac{d_1}{d_1 + d_2} f(C)$$

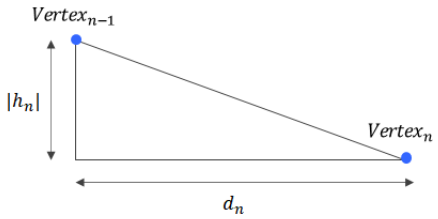
$$f(P) = \frac{d_3}{d_3 + d_4} f(I) + \frac{d_3}{d_3 + d_4} f(J)$$

<Figure 3> Bilinear Interpolation

교통링크가 갖는 종단경사 기울기는 교통링크의 각 Vertex가 갖는 고도차와 해당 Vertex와 인접한 Vertex와의 거리를 이용해서 도출할 수 있다. 예를 들면 $Vertex_n$과 $Vertex_{n-1}$가 서로 인접할 때 $Vertex_n$의 경사도 $slope_n$은 <Figure 4>와 같이 구할 수 있다.

교통노드에 3개 이상의 교통링크가 연결되어 있다면 교차로라고 판단할 수 있다. <Figure 5>의 그림은 4개의 교통링크 $Link_1$, $Link_2$, $Link_3$, $Link_4$가 1개의 교통노드 $Node_1$ 에 연결된 4지 교차로를 나타낸다. $Node_1$을 교통링크 종료지점으로 갖는 $Link_i$와 인접한 $Link_{i+1}$가 이루는 각도를 θ_i라 하였을 때 θ_i의 값은 두 교통류가 이루는 각도이다. 한국건설교통부는 평면교차로 설계 지침에 다음과 같이 설명하였다. 서로 교차하는 교통류는 직각으로 교차하도록 하는 것이 두 교통류의 상대속도를 최소화하고 시

야가 넓어져서 좋다. 즉, 교차각이 작은 경우에는 상충 지점 및 자동차의 회전궤적이 커서 교통사고가 발생하기 쉽다. 따라서 비스듬히 교차하는 형태의 교차로(Y형, X형 등)는 가능한 한 직각에 가깝도록 90도를 기준으로 ±15 이내의 교차로(T형, 십자형)로 설계한다. 본 연구에서는 국내 교차로 기하구조를 고려하여 한국건설교통부가 제시한 평면교차로 설계 지침에 따라 교차로 사잇각 90도를 안전한 시거 확보 기준값으로 사용하였다[8]. 앞에서 정의한 기준을 이용하여 각 θ_i가 90도에서 벗어난 정도를 계산한 값인 $f(\theta_i)$의 평균값 $\mu(f(\theta_i))$을 도출하여 해당 노드가 안정적인 θ_i 값에서 얼마나 벗어났는지 판단할 수 있다. 또한 θ_i 값의 균일정도를 판단하기 위해 $f(\theta_i)$의 변동계수 $CV(f(\theta_i))$을 도출하여 사용하였다.

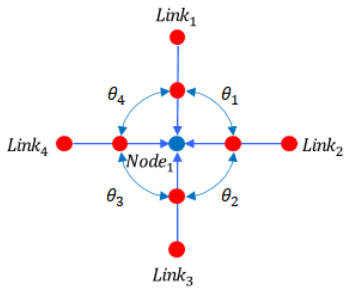


h_n = the altitude value of $Vertex_2$ − the altitude value of $Vertex_1$

d_n= Distance between $Vertex_2$ and $Vertex_1$

$$slope_n = \frac{h_n}{d_n}$$

<Figure 4> Derivation of Vertical Gradient of Traffic Link



$$f(\theta_i) = |90 - \theta_i|$$

$$\mu(f(\theta_i)) = \frac{\sum_{i=1}^{n(Link)} f(\theta_i)}{n(Link)}$$

$$\sigma(f(\theta_i)) = \sqrt{\frac{\sum_{i=1}^{n(Link)} (f(\theta_i) - \mu(f(\theta_i)))^2}{n(Link)}}$$

$$CV(f(\theta_i)) = \frac{\sigma(f(\theta_i))}{\mu(f(\theta_i))}$$

<Figure 5> Analysis of the Angle between Traffic Links with the Same Traffic Node as the Transportation Link End Point

3.3 교통노드 및 교통링크와 교통사고 매핑

본 연구에서는 각 교통노드 및 교통링크에서 1년간 발생하는 교통사고건수를 예측하는 것을 목적으로 한다. 각 교통노드 및 교통링크에서 1년간 발생하는 교통사고 건수 정보를 생성하기 위해 각 교통노드 및 교통링크에 각 교통사고정보를 매핑하였다. 교통노드 기준으로는 교통사고 발생 장소와 150m 거리 이내에 포함된 교통노드 중 가장 근접한 교통노드를 찾아 매핑하였다. 교통링크 기준으로는 교통사고 발생 장소와 1km 거리 이내에 가장 근접한 교통링크에 포함된 2개의 Vertex가 만드는 직선을 찾아 매핑하였다. 이를 통해 각 교통노드 및 교통링크에서 1년간 발생하는 교통사고정보를 생성하였다.

3.4 교통사고 발생당시 환경정보 생성

본 연구에서는 교통사고 발생당시 환경정보를 생성하기 위해 교통소통정보와 기상환경정보를 사용하였다. 교통소통정보는 특정 교통사고가 발생한 교통링크 혹은 교통노드의 일별, 시간별 통행속도를 사용하였다. 교통노드의 통행속도는 교통노드에 연결된 교통링크들의 평균 통행속도 값을 사용하였다. 교통소통정보는 국가교통정보센터에서 제공하는 기준에 맞추어 통행속도별, 도로등급별 정체, 서행, 원활로 구분하여 사용하였다. 이를 통해 교통사고 발생당시 교통소통 상황을 추정하였다. 기상환경정보는 기상청 ASOS 정보를 이용하여 강수여부, 적설여부를 생성하였고 일출 및 일몰 정보를 사용하여 주간 및 야간 여부를 생성하였다. 이를 통해 교통사고 발생당시 기상환경을 추정하였다.

3.5 모델구축용 데이터셋 생성

본 연구에서는 2021년 6월 기준 세종특별자치시에 위치한 교통노드 1,771개, 교통링크 4,769개 정보를 사용하였다. 이중 국가교통정보센터에서 구축한 2020년부터 2021년까지의 교통소통정보를 포함하여 교통노드(1,424개)와 교통링크(3,370개)에서 발생한 2017년부터 2019년까지의 교통사고자료, 교통소통정보, 교통사고 발생당시 날씨, 기하구조자료를 이용하였다. 교통소통정보는 2020년부터 2021년까지의 데이터를 이용하여 교통링크의 요일별과 시간별 평균통행속도를 사용하였다. 교통사고발생당시 날씨정보는 세종특별자치시 데이터의 부재로 인하여 기상청 종관기상관측자료 자료 중에서 가장 가까운 대전광역시(133번) 자료를 사용하였다. 기하구조자료는 좌표데이터와 DEM 데이터를 이용하여 교통노드기반기하구조와 교통링크기반 기하구조정보를 생성하였다. 이를 통해 교통

노드 및 교통링크를 기반으로 특정시점의 기상환경, 교통소통정보를 추정할 수 있는 데이터셋을 생성하였다.

교통노드 기반 교통사고 예측모델에 사용한 데이터셋은 교통노드식별자별, 교통소통정보별, 강수여부별, 적설여부별, 주간여부별로 그룹화하여 생성하였다. 교통노드 기반 교통사고 예측모델에 사용되는 데이터셋 정보는 <Table 1>과 같다.

<Table 1> Traffic Node Based Data Set Information

Year	Number of traffic accidents	Number of observational traffic accidents	Number of data
2017	746	188	34,176
2018	795	229	34,176
2019	922	236	34,176
합계	2,463	653	102,528

교통노드 기반 교통사고 예측모델에 사용되는 데이터셋 구조는 <Table 2>와 같다.

<Table 2> Traffic Node Based Data Set Structure

Feature	Explanation
NODE_ID	Node ID
DAY	1:Day , 0:Night
RAIN	Precipitation
SNOW	Snow cover
TF_JAM	Traffic congestion
TF_SLOW	Slow traffic speed
TF_SMOOTH	Smooth traffic speed
NT_101	Intersection
NT_102	Road star and point
NT_103	Property change point
NT_104	Whether road facilities
NT_105	Administrative boundary
NT_106	Connecting part
NT_108	IC and JC
TURN_P	Rotation limit
LINK_NUM	Number of connected links
IN_LINK_NUM	Number of incoming links
OUT_LINK_NUM	Number of outgoing links
LANES_VALUE	Number of lanes
IN_LANES_SUM	Number of incoming lanes
OUT_LANES_SUM	Number of outgoing lanes
SPD_MEAN	Average speed limit
SPD_CV	Coefficient of variation of speed limit
UTURN	Number of U-turns allowed
LEFT	Number of left turns prohibited
STRAIGHT	Number of straight-through prohibited
RIGHT	Number of right turns prohibited
LANES_MEAN	Average number of lanes on each link
LANES_CV	Coefficient of variation in the number of lanes on each link
DEGREE_MEAN	$\mu(f(\theta_i))$
DEGREE_CV	$CV(f(\theta_i))$
STRAIGHT_MEAN	Average straightness of the incoming link
STRAIGHT_CV	Coefficient of variation of the straightness of the incoming link
SLOPE_MEAN	Average number of downhill links entering
SLOPE_CV	Coefficient of variation in the number of downhill links entering
CURVE_CNT_MEAN	Average number of curve sections of incoming link
CURVE_CNT_CV	Coefficient of variation of the number of curved sections of the incoming link
ACC_CNT	Number of traffic accidents per year

교통링크 기반 교통사고 예측모델에 사용한 데이터셋은 교통링크식별자별, 교통소통정보별, 강수여부별, 적설여부별, 주간여부별로 그룹화하여 생성하였다. 교통링크 기반 교통사고 예측모델에 사용되는 데이터셋 정보는 <Table 3>과 같다.

<Table 3> Traffic Link Based Data Set Information

Year	Number of traffic accidents	Number of observational traffic accidents	Number of data
2017	746	670	79,320
2018	795	695	79,320
2019	922	766	79,320
합계	2,463	2,131	237,960

교통링크 기반 교통사고 예측모델에 사용되는 데이터셋 구조는 <Table 4>와 같다.

<Table 4> Traffic Link Based Data Set Structure

Feature	Explanation
LINK_ID	Link ID
DAY	1:Day , 0:Night
RAIN	Precipitation
SNOW	Snow cover
JAM	Traffic congestion
SLOW	Slow traffic speed
SMOOTH	Smooth traffic speed
LANES	Number of lanes
MAX_SPEED	Speed limit
LENGTH	Road length

Feature	Explanation
UP_SLOPE_MEAN	Average slope uphill
UP_SLOPE_CV	Coefficient of variation of uphill slope
DOWN_SLOPE_MEAN	Average slope downhill
DOWN_SLOPE_CV	Coefficient of variation of downhill slope
SLOPE_LENGHT_RATE	Ratio of sections with slope
CURVE_SET_CNT	Number of curve sections
CURVE_MEAN	Average curvature
CURVE_CV	Coefficient of variation of curvature
CURVE_LENGTH_RATE	Ratio of curved section
RR_101	High-speed national highway
RR_102	City highway
RR_103	General national road
RR_104	Special Metropolitan City
RR_105	National support map
RR_106	Province
RR_107	City and county
RT_0	General road
RT_1	Overpass
RT_2	Underpass
RT_3	Bridge
RT_4	Tunnel
CURVE_SET_RATE	Vertex ratio satisfying CURVE_SET_CNT>0 and >1
ACC_CNT	Number of traffic accidents per year

4. 예측모델 구축 및 모델평가

4.1 예측모델 구축

교통노드 기반 교통사고 예측모델과 교통링크 기반 교통사고 예측모델의 데이터셋은 2017년, 2018년 데이터는 병합 후 8:2 비율로 학습용 및 검증용 데이터셋을 생성하였으며, 2019년 데이터셋은 평가용 데이터셋으로 사용하였다.

교통노드 기반 교통사고 예측모델은 교통노드 기반 교통사고 예측모델 데이터셋에서 1년간 발생한 교통사고건수(ACC_CNT)를 종속변수로 사용하였으며, 교통노드식별자(NODE_ID)를 제외한 나머지 37개의 변수를 독립변수로 사용하여 XGBoost 모델에 적용한다. 교통링크 기반 교통사고 예측모델도 마찬가지로 1년간 발생한 교통사고건수를 종속변수로 사용하였으며, 교통링크식별자(LINK_ID)를 제외한 나머지 31개의 변수를 독립변수로 사용하여 XGBoost 모델에 적용한다. 교통노드 기반 교통사고 예측모델과 교통링크 기반 교통사고 예측모델은 다음과 같은 과정을 거쳐 모델을 생성한다. 먼저 학습용 데이터셋을 통해 모델을 생성하고, 검증용 데이터셋을 통해 모델의 성능을 확인하였다. XGBoost 알고리즘에서의 하이퍼 파라미터(Hyper Parameter)를 변경하여 다시 재학습하는 과정을 거치며 가장 좋은 성능을 갖는 모델을 선정하였다. 이에 대한 일반적인 튜닝은 Grid Search, Random Search, Baysian Optimization 방식을 사용한다. 본 연구에서는 Grid Search 방식을 사용하여 최적의 하이퍼 파라미터를 결정하였다. 교통노드 기반 교통사고 예측모델에서의 하이퍼 파라미터는 각 트리(스탭)마다 사용할 칼럼의 비율(colsample_bytree)을 0.7로, 학습률(learning_rate)은 0.01로, 최대 트리의 깊이(max_depth)는 6으로, 각각의 트리에서 가지추가를 위한 최소 사례 수(min_child_weight)는 5로, 목적함수(objective)는 'reg:squarederror'로, 각 스탭마다 사용할 샘플 비율(subsample)은 1로 각각 설정하였다. 반면에, 교통링크 기반 교통사고 예측모델의 경우에는 각 트리(스탭)마다 사용할 칼럼 비율(colsample_bytree)은 0.6, 학습률(learning_rate)은 0.01로, 최대 트리의 깊이(max_depth)는 8로, 그리고 각각의 트리에서 가지 추가를 위한 최소 사례 수(min_child_weight)는 4로, 목적 함수(objective)는 'reg:squarederror'로, 각 스탭마다 사용할 샘플 비율(subsample)은 1로 설정하였다(<Table 5> 참조).

<Table 5> Traffic Node and Link Based Model Hyper Parameters

Hyper Parameter	Value	
	Traffic node based model	Traffic link based model
colsample_bytree	0.7	0.6
learning_rate	0.01	0.01
max_depth	6	8
min_child_weight	5	4
objective	reg:squarederror	reg:squarederror
subsample	1	1

4.2 모델평가

XGBoost를 사용하는 분류모델은 일반적으로 F1 스코어를 평가지표로 사용하지만 본 연구에서 사용된 XGBoost 모델은 교통노드/교통링크별, 교통소통정보별, 기상환경별 1년간 발생하는 사고건수를 예측하는 회귀모델이기 때문에 평균제곱근오차(Root Mean Square Error, RMSE)를 평가지표로 사용하였다. 교통노드 기반 교통사고 예측모델은 2019년 교통노드 기반 교통사고 예측모델 데이터셋을 사용하였고 예측모델에 대한 평가결과는 <Table 6>과 같다.

교통링크 기반 교통사고 예측모델은 2019년 교통링크 기반 교통사고 예측모델 데이터셋을 사용하여 예측모델을 평가하였고 평가결과는 <Table 7>과 같다.

<Table 6> Traffic Node Based Model Evaluation Result

Evaluation Standard	Training Data Set	Test Data Set
	Value	Value
MAE	0.1912	0.1910
Maximum absolute error	2.8	3.7992
Minimum absolute error	0.1826	0.1826
RMSE	0.2036	0.2035
Standard deviation of absolute error	0.0698	0.0702

<Table 7> Traffic Link Based Model Evaluation Result

Evaluation Standard	Training Data Set	Test Data Set
	Value	Value
MAE	0.1937	0.1945
Maximum absolute error	3.7637	4.799
Minimum absolute error	0.1824	0.1238
RMSE	0.2084	0.2107
Standard deviation of absolute error	0.0769	0.0811

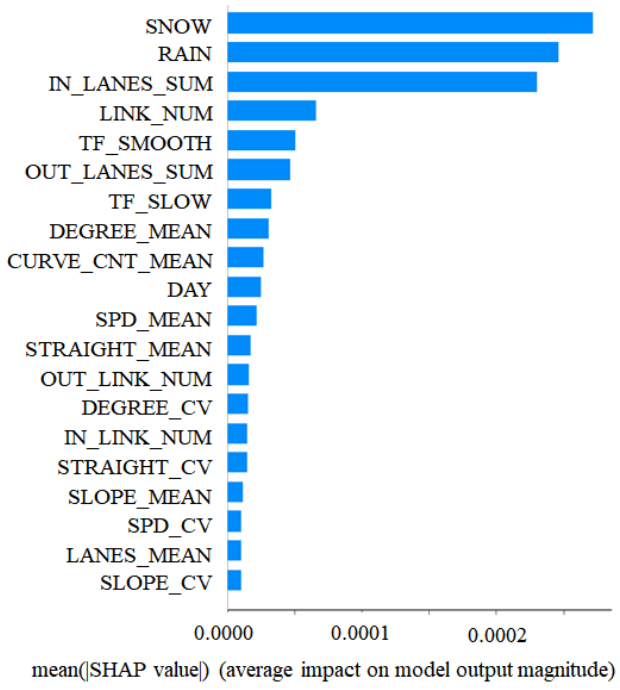
본 연구에서 개발한 교통노드 기반 교통사고 예측모델과 교통링크 기반 교통사고 예측모델의 상대적인 평가를 위해 선행연구인 '딥러닝을 이용한 고속도로 교통사고 예측모델 개발(Traffic Accident Prediction Model Using Deep Learning)'[16] 및 '확률모수를 이용한 교통사고 예측모형 개발(Accident Frequency Estimation Models Using Random Parameter)'[3]과의 성능 비교를 진행하였다. 앞에서 언급한 '딥러닝을 이용한 고속도로 교통사고 예측모델'은 781개의 관측수를 갖고 확률모수를 이용한 교통사고 예측모형은 57개의 관측수를 갖는다. 반면에 본 연구에서의 교통노드 기반 교통사고 예측모델은 관측수가 34,176개, 교통링크 기반 교통사고 예측모델의 관측수는 79,320개로 선행연구에 비해 관측수가 많다. 다만 전체 관측대상 중 99% 이상이 교통사고가 발생하지 않은 관측대상이기 때문에 평균제곱근오차(RMSE)가 상대적으로 낮게 측정 된다. 이런 이유로 교통노드 기반 교통사고 예측모델과 교통링크 기반 교통사고 예측모델에서는 교통사고 발생건수가 1 이상인 관측치를 사용하여 평균절대편차(Mean Absolute Deviation, MAD)와 평균제곱근오차(RMSE)를 도출하였다. 또한, '확률모수를 이용한 교통사고 예측모형 개발' 연구에서는 평균제곱근오차(RMSE)를 지표로 사용하지 않았지만 객관적인 성능비교를 위해 논문의 "모형을 적용한 도로 안정성 평가결과" 내용 중 확률적 모수모형(RPM) 데이터 기반으로 평균제곱근오차(RMSE)를 별도 계산하여 비교하였다. 성능비교 결과는 <Table 8>과 같다.

<Table 8> Predictive Model Relative Evaluation Result

Model	MAD	RMSE
Traffic Accident Prediction Model Using Deep Learning	2.52	3.43
Accident Frequency Estimation Models Using Random Parameter	1.11	1.865
Traffic link based traffic accident prediction model	0.03	1.05
Traffic node based traffic accident prediction model	0.03	1.09

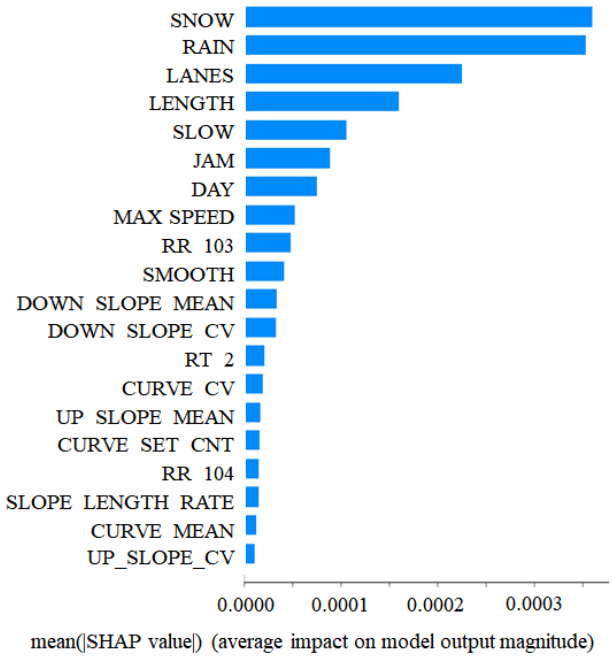
성능비교 결과 교통노드 기반 교통사고 예측모델과 교통링크 기반 교통사고 예측모델 모두 선행연구인 '딥러닝을 이용한 고속도로 교통사고 예측모델'과 '확률모수를 이용한 교통사고건수 예측모형'의 평균절대편차(Mean Absolute Deviation, MAD)와 평균제곱근오차(RMSE) 지표보다 우수한 지표를 보인다. 다만, 선행연구에서 개발한 모델별로 사용하는 데이터가 다르고 실제 교통사고 발생위치를 특정하는 지역의 특성이 다르기 때문에 어떤 모델이 더 우수하다고 판단하기 어렵다. 하지만 본 평가를 통해 선행연구에서 개발한 교통사고 예측모델과 비교하여 교통노드 기반 교통사고 예측모델과 교통링크 기반 교통사고 예측모델의 성능이 낮지 않다는 것을 알 수 있다.

교통노드 기반 교통사고 예측모델의 중요 변수를 SHAP 방법을 이용하여 도출한 결과는 <Figure 6>과 같다.



<Figure 6> Traffic Node Based Model SHAP Variable Importance

교통링크 기반 교통사고 예측모델의 중요 변수를 SHAP 방법을 이용하여 도출한 결과는 <Figure 7>과 같다.



<Figure 7> Traffic Link Based Model SHAP Variable Importance

교통노드 기반 교통사고 예측모델의 중요도 상위 10개 변수는 적설여부, 강수여부, 진입하는 차선 개수, 연결된 링크 개수, 서행여부, 진출하는 차선 개수, 정체 여부, 평균 교차로 사잇각, 진입하는 링크의 평균 곡선개수, 주간 여부 순으로 나타났다. 교통링크 기반 교통사고 예측모델의 중요도 상위 10개 변수는 적설여부, 강수여부, 차로수, 도로 길이, 서행 여부, 정체 여부, 주간 여부, 제한 속도, 일반국도 여부 순으로 나타났다.

5. 결 론

본 연구에서는 교통노드 및 교통링크 기반으로 1년간 발생하는 교통사고건수 예측을 위해 XGBoost를 이용한 교통노드 기반 교통사고 예측모델과 교통링크 기반 교통사고 예측모델을 제안하였다. 먼저 자료구축을 위해 교통사고정보, 교통노드정보, 교통링크정보, 교통노드 회전정보, DEM 정보, 세종특별자치시 교통소통정보, 대전광역시 기상정보를 수집하였다. 수집한 데이터를 기반으로 교통노드 및 교통링크의 기하구조를 생성하고 교통사고정보를 교통노드 및 교통링크와 매핑하여 교통사고 발생당시 환경정보를 생성하고, 모델구축용 데이터셋 생성을 진행하였다.

교통노드 기반 교통사고 예측모델 데이터셋과 교통링크 기반 교통사고 예측모델 데이터셋을 각각 학습용, 검증용, 평가용으로 나누었고, XGBoost를 적용하여 모델학습 및 모델평가를 진행하였다. 교통노드 기반 교통사고와 교통링크 기반 교통사고 예측모델의 RMSE는 각각 0.2035와 0.2107로 나타났다.

SHAP 방법을 활용하여 교통사고 발생건수에 영향을 주는 변수를 선정하였다. 교통노드 기반 교통사고 예측모델의 주요변수 5개는 적설여부, 강수여부, 진입하는 차선 개수, 연결된 링크 개수, 서행여부이며, 교통링크 기반 교통사고 예측모델의 주요변수 5개는 적설여부, 강수여부, 차로수, 도로 길이, 서행 여부로 나타났다. 날씨, 교통소통상태, 노출도로면적이 교통사고 발생건수가 가장 큰 영향을 주는 것을 확인할 수 있었다.

이와 같은 결과를 통해 다음과 같이 해석할 수 있다. 운전자는 눈과 비가 내리는 날씨, 교차각이 작은 교차로 통과 시 시거 확보, 통과속도가 느린 교차로 진입 시 이전 구간에서 서행 및 안전거리 확보, 서행 및 정체 구간 진입 시 서행, 서행 및 정체 구간 내에서 안전거리 확보에 주의해야 하며 지자체는 교차각이 작은 교차로의 사고방지 및 상습 서행구간 진입 구간에 교통사고 방지를 위한 대비 마련이 필요하다.

6. 연구의 한계점 및 향후 연구

본 연구에서는 교통노드 및 교통링크 기반 교통사고 예측모델에 세종특별자치시 데이터만 사용하였지만 교통노드 및 교통링크가 구축되어 있는 모든 지역에 본 연구에서 제안한 예측모델을 적용할 수 있다. 또한 실시간으로 기상정보와 교통소통정보를 이용하여 5분 주기로 새로운 데이터셋을 생성하고 특정 시점에 도로별 상대적인 교통사고 위험도를 추정하여 <Figure 8>과 같이 실시간 사고위험 지역을 시각화 할 수 있다. 이를 통해 교통사고 주의 구역 선정, 교통안전시설물 및 교통신호체계 정비대상 우선순위 도출, 교통사고위험지역 실시간 모니터링 등에 활용할 수 있을 것으로 기대한다.

본 연구에서 사용된 데이터셋에는 교통사고 발생건수가 0건인 데이터가 대부분을 차지한다. 따라서 데이터 불균형 문제를 갖고 있으며, 이런 문제는 오버샘플링 또는 언더샘플링으로 해결할 수 없었다. 기존의 방법보다 우수한 샘플링 방법을 적용한다면 우리의 연구보다 우수한 성능의 모델을 개발할 수 있을 것으로 기대한다. 또한 교통사고는 도로의 기하구조, 기상환경, 교통소통 상황뿐만 아니라 운전자의 심리상태, 시거문제 등 복합적인 문제로 발생할 수 있

다. 향후 운전자의 심리상태, 시거상태, 교통시설물설치정보 등을 반영한 데이터셋을 활용한다면 보다 더 우수한 성능의 모델을 개발할 수 있을 것으로 기대한다.

Hazard class	1	2	3	4
Color	Green	Yellow	Orange	Red

<Figure 8> Real-time Visualization of Accident Risk Areas in Sejoing City

Acknowledgement

This work is supported by VAIV Company Inc. which carried out the Sejong Technopark Foundation's 「Sejong Autonomous Driving Big Data Control Center Construction and Operation」 Project.

References

[1] Chen, T. and Guestrin, C., XGBoost: A Scalable Tree Boosting System, KDD'16, 2016, pp. 785-794.

[2] Cho, N.H., Jun, C.M., and Kang, Y.O., A visualization of traffic accidents hotspot along the road network, *Journal of Cadastre & Land InformatiX*, 2018, Vol.48, No.1, pp. 201-213.

[3] Im, J.B., Development of Accident Frequency Estimation Models Using Random Parameter, University of Seoul, 2015.

[4] Jeong, Y.H. and Choi, Y.W., A study on the analysis of urban highways traffic accident's impact factors based on building discriminant models: Busan Metropolitan City, *KSCE Journal of Civil and Environmental Engineering Research*, 2014, Vol. 34, No. 4, pp. 1269-1278.

[5] Kang, Y.G., Traffic Accident Frequency Prediction Model in Urban Signalized Intersections with Intelligent Theories, University of Seoul, 2008.

[6] Kang, Y.O., Son, S.R., and Cho, N.H., Analysis of traffic accidents injury severity in seoul using decision trees and spatiotemporal data visualization, *Journal of Cadastre & Land InformatiX*, 2017, Vol. 47, No. 2, pp.233-254.

[7] Korea Meteorological Agency Data Open Portal, http://data.kma.go.kr.

[8] Korea Ministry of Construction and Transportation, Plan Intersection Design Guidelines, 2004.

[9] Korea National Transport Information Center, http://its.go.kr.

[10] Korea Traffic Accident Analysis System, http://taas.koroad.or.kr.

[11] Korean National Police Agency, http:// police.go.kr.

[12] Lee, K.J., Jung, I.G,. Noh, Y.H., Yoon, S.G., and Cho, Y.S., The effect of road weather factors on traffic accident: Focused on Busan area, *Journal of the Korean Data and Information Science Society*, 2015, Vol. 26, No. 3, pp. 661-668.

[13] Lee, S.H., Park, M.H., and Woo, Y.H., A study on developing crash prediction model for urban intersections considering random effects, *The Journal of The Korea Institute of Intelligent Transport Systems*, 2015, Vol.14, No.1, pp. 85-93.

[14] Lundberg, S.M. and Lee, S.I., A Unified Approach to Interpreting Model Predictions, Advances in Neural Information Processing Systems, 2017, 30.

[15] Oh, H. U., Correlation between design consistency and accident rates based on standard deviations of highway design elements, *International Journal of Highway Engineering*, Vol.11, No.2, 2009, pp. 159-166.

[16] Park, M.H., Relationship between interstate highway accidents and heterogeneous geometrics by random parameter negative binomial model: A case of interstate highway in Washington State, USA, *J. Korea Soc. of Civil Eng.*, 2013, Vol. 33, No. 6, pp. 2437-2445.

[17] Ryu, J.D., Development of Expressway Traffic Accident Prediction Model Using Deep Learning, University of Ajou, 2018.

[18] United States Geological Survey, https://earthexplorer.usgs.gov.

ORCID

Un Sik Kim | http://orcid.org/0000-0002-7465-0298
Young Gyu Kim | http://orcid.org/0000-0003-2793-2206
Joong Hoon Ko | http://orcid.org/0000-0001-8168-9126