머신러닝을 이용한 교통사고 사상자 수 예측 :서울시 공공데이터를 대상으로

남명우[O], 박두서[*], 장영준[*], 이홍철[*]
[O]고려대학교 산업경영공학과,
[*]고려대학교 산업경영공학과
e-mail: {ds020902, yjang11, hclee}@korea.ac.kr[*], nmw0221@korea.ac.kr[O]

Prediction Of Traffic Accident Casualties Using Machine Learning : For Seoul Public Data

Myung-woo Nam[O], Doo-Seo Park[*], Young-Jun Jang[*], Hong-Chul Lee[*]
[O]Dept. of Industrial Management Engineering, Korea University,
[*]Dept. of Industrial Management Engineering, Korea University

● 요 약 ●

경제 성장과 함께 자동차의 수요가 늘어남에 따라 교통사고 발생 빈도는 꾸준히 증가하고 있다. 이에, 본 연구에서는 교통사고를 야기하는 도로 및 기상환경과 같은 조건을 활용하여 기계학습 모델을 통해 서울시 교통사고 사상자 수를 예측하는 모형을 찾고자 한다. 활용한 데이터는 도로교통 공단에서 제공하는 교통사고 사상자 수 정보를 포함하는 데이터로 2015년부터 2018년도까지 데이터를 학습에 사용하였고 2019년도 데이터를 테스트 평가에 사용하였다. 실증연구를 통해 트리 기반의 모델 별 성능을 비교하였으며 본 연구에 대한 결과는 사고 발생 시 우선순위에 의한 구조활동이 가능하게 함과 도로상황 및 기상을 고려한 안전운전 가이드 지식으로 활용될 수 있다.

키워드: 머신러닝, 교통사고, 사상자수 예측, DecisionTree, RandomForest, LightGBM, XGBoost

I. Introduction

대한민국은 1980년 이후 국민경제의 성장으로 편리한 교통수단으로서의 자동차의 수요가 증가하였고 도로교통 여건이 발전하고 있다. 하지만, 자동차 수의 증가는 환경문제, 교통사고 및 교통 문제 증가 등을 유발하게 되었다. 2000년도 들어서서는 사망자 수가 최근 몇 년간 감소하였지만, 연평균 사고 건수가 증가 추세로 2019년도 사건 건수는 전년 대비 3.5% 증가하였다[1]. 이는 차량의 발달에 따라 탑승자 안전이 확보되어 사망자 수는 감소하였으나 아직 교통사고를 미연에 방지할 수 있는 대책은 부족하다고 생각할 수 있다. 이에 따라, 교통사고의 사상자 수를 정확히 예측하고 동시다발에 발생하는 교통사고에 따른 다양한 속성 및 원인 분석이 필요하며 신속한 판단이 요구된다. 따라서, 본 논문에서는 여러 가지 상황 속에 발생할 수 있는 교통사고의 요인들을 고려하여 기계학습을 이용한 사상자 수 예측하고 이에 따른 최적의 모델을 찾고자 한다. 본 연구에 사용한 알고리즘은 DecisionTree 모델과 대표적인 앙상블 모형인 RandomForeset, XGBoost 그리고 LightGBM을 사용하였다. 연구에 대한 결과는 동시다발적인 교통사고 발생 시 우선순위에 의한 구조 활동 시스템 정착에 도움을 줄 것으로 예상한다.

II. Preliminaries

1. Related works

최근 들어 국내에서 꾸준한 교통사고의 증가로 인하여 많은 문제가 대두되고 있다. 이에 따라, 교통사 고의 발생원인과 특성을 밝히는 것이 우선시 되어야 한다[1]. 관련하여, 많은 교통사고와 관련된 교통사고 예측 연구가 진행되고 있으며 기존 연구에서는 교통사고 시도별 특성을 분석하고 도로교통사고 특성을 설명하는 거시적 차원의 사고예측모형을 개발하여 회귀분석 모형과 시계열분석 모형으로 나누어 제시하였다[2].

[3]은 국내 TAAS(Traffic Accident Analysis System)에서 제공

하는 교통사고 데이터를 활용하여 상해 심각도를 6단계로 분류 예측하였다.

본 연구에서는 교통사고 발생 변수를 고려하여 교통 사고의 사상자 수를 예측하고자 한다.

III. The Proposed Scheme

1. 데이터 셋 구축

본 연구에서의 데이터는 도로교통공단에서 제공하는 2015년부터 2019년도까지 총 1,571개의 교통사고 데이터와 기상 자료 개방 포털에서 제공하는 해당 기간의 서울시 기상 데이터를 사용하였다. 수집된 데이터 이후 실험에 사용된 데이터의 형식과 형태는 가공되었으며 내용은 테이블 1과 같다.

Table 1. 수집 데이터셋

변수명	설명
발생년월일시분	사고 발생 일자 및 시간
사고유형	사고 유형 (충돌, 추돌, 보행, 기타 등)
주·야	사고 발생 시 주·야간 구분
도로형태	사고 발생지역 도로형태 (터널안, 횡단보도부근, 교량위, 주차장 등)
법규위반	가해자 법규위반 사항 (과속, 신호위반, 중앙선 침범등)
발생지시군구	사고 발생지역 (구)
기 온	공기의 온도 (단위 : 섭씨)
풍 속	바람의 세기 (단위 : m/s)
지면온도	지면의 온도
강수량	일정 지역에 내린 물의 총량 (단위 : mm)
시 정	사방을 둘러보아 가장 짧은 거리 (단위 : m)
적 설	일정 지역에 내린 눈의 총량 (단위 : cm)
사상자수	사고에 따른 사상자수 (사망자수+중상수+경상자수+부상자수)

2. 방법론

2.1 앙상블 모델

앙상블 모델이란 동일한 학습 알고리즘을 사용해서 여러 개의 모형으로 나온 예측 결과를 다수결 법칙 또는 평균을 이용해 통합하여 최종적인 의사결정에 활용하는 방법으로 Variance와 Bias를 감소시킨다. 앙상블 모델에는 부스팅(boosting), 배깅(bagging) 방식이 있으며 Boosting은 정답과 오답에 대한 가중치를 다르게 하여 어려운 문제를 좀더 집중하는 반면, Bagging은 랜덤하게 복원추출 후 결과를 집계하는 하는 차이점이 있다[4]. 앙상블 모델에는 AdaBoost, CatBoost, GradientBoost 등 다양한 모델이 있지만 본 논문에서는 대표적인 XGBoost와 LightGBM을 사용하였다.

2.2 RandomForest

RandomForest는 DecisionTree의 단점을 보완하고자 Breiman에 의해 모형이 개발되었다[5]. RandomForest는 데이터를 부트스트랩(bootstrap) 하여 포레스트를 구성 한다. 전체 데이터를 전부 이용하는 것이 아닌 분할된 노드에서 랜덤으로 선택된 샘플을 이용하여 트리를 생성하게 된다. 무작위로 노드에 대한 모든 변수를 고려 하는 것이 아닌 랜덤으로 선택된 샘플에 대해서 배깅 과의 차이가 있으며 무작위로 변수들을 선택 과정이 반복되게 된다. 모든 변수 중 최적의 변수를 선택하는 것이 아닌 변수 중 일부분만 선택하고 그 일부 중에서 최적의 변수를 선택하게 되고 이러한 방식으로 예측 값이 다양하게 되어 앙상블 효과를 가지고 오게 된다.

2.3 XGBoost

XGBoost는 병렬처리와 최적화를 장점으로 내세우는 graident boosting machine(GBM)의 한 방법으로서 트리기반의 모델에 잔차(residual)를 학습시키는 부스팅 (boosting) 기법을 적용한 머신러닝 모델이다. GBM의 단점을 보완하여 각종 머신러닝 경쟁에서 다른 기존 모델에 비해 속도나 성능 면에서 우수함을 보여주었다. 또한 많은 하이퍼파라미터를 지원하여 다양한 데이터에 대한 유연한 학습을 가능하게 해주며 CART(Classification And Regression Rree) 방식을 이용하여 생성된 트리모 델과 리프(leaf)의 우위비교로 입력 변수들 간의 중요도를 확인할 수 있다.

2.4 LightGBM

LightGBM은 기존의 GBM 계열과 다르게 leaf-wise 방식을 이용하여 더욱 복잡한 모형을 만들어 우수한 정확성을 보여준다. 빠른 훈련처리 속도라는 장점을 가지고 있지만 쉽게 과적합(overfitting)이 되는 단점을 가지고 있어 충분한 데이터가 있을 때 적합한 모형이다.

2.5 하이퍼 파라미터 최적화 과정

머신러닝 모델 설정에 있어서 입력변수들이 결정되고 데이터의 특성에 따라 하이퍼 파라미터의 조절이 필수 적이다. 하이퍼 파라미터는 이론적으로 정하는 것이 아닌 경험적으로 사람이 직접 데이터에 특성에 따라 설정 되어야 하는 값으로 최적의 하이퍼 파라미터를 찾는 것은 높은 성능을 얻기 위한 필수적인 작업이다. 본 논문 에서는 데이터에 대한 최적의 하이퍼파라미터 조합을 찾아 주고 모델의 일반화 성능을 최대로 높여주기 위해 그리드 서치(Grid Search)를 수행하였다.

표 2는 5-fold Cross Validation을 통한 그리드 서치 이후 RandomForest의 최적 하이퍼 파라미터로의 결과 로서 생성할 나무의 개수를 의미하는 n_estimators와 나무의 깊이를 뜻하는 max_dept는 다음과 같다.

Table 2. RandomForest 최적 하이퍼파라미터

하이퍼파라미터	RandomForest
CV	5
n_estimator	500
max_depth	2

Table 3. LightGBM, XGBoost 최적 하이퍼파라미터

하이퍼파라미터	LightGBM	XGBoost
colsample_bytree	0.5	0.5
subsample	0.2	0.8
num_leaves	10	-
n_estimators	200	500
learning_rate	0.01	0.01
max_depth	-	2

표 3은 LightGBM과 XGBoost의 그리드서치를 통한 최적의 하이퍼 파라미터의 결과 값이다. 그리드서치는 시간이 오래 걸리는 단점이 있지만 주어진 범위 내에서 가장 좋은 결과를 얻을 수 있다.

3. Experimental Result

표 4를 보면 사상자 수 예측에 따른 머신러닝 모델간 성능을 MAPE, RMSE의 지표로 비교한 결과이다.

DecisionTree에 비해 앙상블 모델이 성능이 더 좋은 결과를 보여주는 것을 확인하였고, 실제값과 예측값의 예측오차 비율을 보여주는 MAPE 지표에서 4개의 모델중 LightGBM 모델에서 가장 낮은 값을 보여주었다.

Table 4. 모델간 성능비교

ML model	MAPE	RMSE
DecisionTree	59.3	2.76
RandomForest	29.3	2.45
XGBoost	29.0	2.45
LightGBM	**27.8**	**2.45**

4. 변수중요도

아래 Fig. 1은 모델별 변수중요도를 나타냈으며, 이를 통해 모델별 변수들이 서로 다르게 중요도를 가지는 것을 알 수 있다. 4개 모델 공통적으로 사고유형, 법규 위반, 지면 온도, 시정거리가 높은 변수 중요도를 가지는 것을 확인할 수 있다. 하지만 변수중요도에서 상위권에 있지 않은 변수들은 특성이 유용하지 않다는 뜻은 아니다. 모델에 따라 학습에 중요한 영향을 미치는 변수를 다르게 선택한 것이며, 다른 변수가 유사한 정보를 가지고 있어 상대적으로 중요하지 않은 변수로 나타날 수 있다.

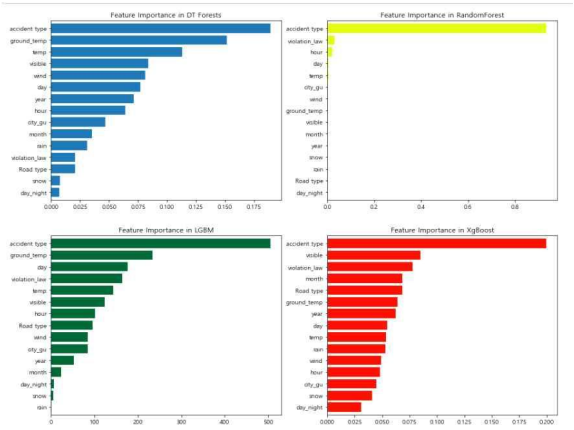

Fig. 1. 모델별 변수중요도

IV. Conclusions

본 연구에서는 도로교통 공단에서 제공하는 사망 교통사고에 대한 데이터를 사용하여 DecisionTree 모델과 트리 기반의 앙상블 모형인 RandomForest, XGBoost, LightGBM 모형을 이용하여 19년도 교통 사고 사상자 수 대해 예측을 하였다. RMSE, MAPE 평가지표를 이용하여 모델별 성능 비교를 하며 교통 사고 사상자 수 예측과정에서 결과 변화를 살펴보고 평가모형을 비교분석 하였다. 본 연구의 한계점으로 는, 공공데이터에서 얻은 교통사고 데이터는 사망자가 발생한 데이터에 한정되어있어 사망자가 발생하지 않은 경우는 고려하지 못하는 점과 데이터의 사망자 수는 실제로 1~2명이거나 대형사고일 경우 수십 명에 달하는 경우가 있어 중간값이 많지 않아 예측정확도가 떨어질 수 있다. 머신러닝 기법 중 트리 기반 앙상블 모델만을 사용하였으며 향후 연구에서는 추가 연도의 데이터를 확보하여 전국적 사고 예측을 수행 하고자 하며 예측 확률을 높이기 위해서 각 사고 발생 주변 시설, 인구분포, 사상자들의 연령대나 성별등 다양한 요소들과 통합되어 입력변수들의 추가하 고, 또 다른 머신러닝 및 딥러닝 기법들을 적용하는 통합적인 연구가 필요하다고 생각된다.

ACKNOWLEDGEMENT

본 논문은 BK21 플러스 사업(고려대학교)으로 지원된 연구임.

REFERENCES

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

[2] Sang-Jin Han. (2007). Road Accident Characteristics in Metropolitan Cities and Provinces. Journal of Environmental Studies, 46(), 211-220.

[3] Yongbeom Lee, Eungi Cho, Changyong Yoon, Seongkeun Park. (2019). A Study on Prediction of Passenger's Injury

Grade Using Public Traffic Accident Database and Machine Learning. The transactions of The Korean Institute of Electrical Engineers, 68(7), 866-871.

[4] Breiman, L. (2001). RandomForests. Machine learning 45 (1), 5–32.