Journal of Korea Society of Industrial Information Systems Vol.28 No.6, Dec. 2023 :11-20
http://dx.doi.org/10.9723/jksiis.2023.28.6.011 ISSN:1229-3741

자동차 사고 경상환자의 장기입원 예측 모델 개발+

(Development of Long-Term Hospitalization Prediction Model for Minor Automobile Accident Patients)

이 덕 규[1], 남 동 현[2]*, 허 성 필[3]*
(DoegGyu Lee, DongHyun Nam, and Sung-Phil Heo)

요 약 자동차보험 교통사고 진료비는 매년 증가하고 있다. 본 연구는 교통사고 진료비용 상승의 주요 항목인 경상환자 중 장기입원환자(18일 이상)를 예측하는 모델을 decision tree 등 5개 알고리즘을 이용하여 생성하고, 장기입원에 영향을 미치는 요인을 분석했다. 그 결과, 예측 모델의 정확도는 91.377 ~ 91.451이며 각 모델 사이에 큰 차이점은 없었으나 random forest와 XGBoost 모델이 91.451로 가장 높았다. 설명변수 중요도에 있어서 병원 소재지, 상병명, 병원 종류 등 장기환자군과 비 장기입원 환자군 사이에 모델마다 상당한 차이가 있었다. 모델 평가는 훈련 데이터의 교차검증(10회)한 모델별 평균 정확도와 실험 데이터의 정확도를 상호 비교한 결과로 검정했다. 설명변수 유의성 검증을 위해 범주형 변수는 카이제곱 테스트를 실시하였다. 본 논문의 연구 결과는 경상 환자들의 과잉진료 및 사회적 보험료 비용을 줄이는 진료행태 분석에 도움이 될 것이다.

핵심주제어: 머신러닝, 자동차사고, 경상환자, 장기입원환자, 예측모델

Abstract The cost of medical treatment for motor vehicle accidents is increasing every year. In this study, we created a model to predict long-term hospitalization(more than 18 days) among minor patients, which is the main item of increasing traffic accident medical expenses, using five algorithms such as decision tree, and analyzed the factors affecting long-term hospitalization. As a result, the accuracy of the prediction models ranged from 91.377 to 91.451, and there was no significant difference between each model, but the random forest and XGBoost models had the highest accuracy of 91.451. There were significant differences between models in the importance of explanatory variables, such as hospital location, name of disease, and type of hospital, between the long-stay and non-long-stay groups. Model validation was tested by comparing the average accuracy of each model cross-validated(10 times) on the training data with the accuracy of the validation data. To test of the explanatory variables, the chi-square test was used for categorical variables.

Keywords: Machine Learning, Auto Accident, Minor Injuries, Long Term Care, Predictive Models

* Corresponding Author: omdnam@sangji.ac.kr, spheo@gwnu.ac.kr
(These authors contributed equally to this work)
+ 이 논문은 한국연구재단 바이오·의료기술개발사업(NRF-2022M3A9E4017033) 및 2019년도 강릉원주대학교 신임교원 연구비 지원에 의하여 수행한 연구임.

Manuscript received August 25, 2023 / revised September 27, 2023 / accepted November 19, 2023
1) 숭실대학교 IT정책경영학과, 제1저자
2) 상지대학교, 교신저자
3) 강릉원주대학교, 교신저자

1. 서 론

2023년 발표한 도로교통공단 교통분석시스템에 따르면 연도별 자동차 사고 건수는 2018년도 217,148건에서 2022년 196,836건으로 점진적으로 감소하였다(KoROAD. 2023). 반면에 교통사고 진료비는 2018년 1조 9,762억 원에서 2022년 2조 5,142억 원으로 매년 증가하고 있다(HIRA, 2022). 자동차보험은 국민들이 교통사고 시 치료를 보장하게 하는 국가 사회적 의무보험이다. 국토교통부는 경상환자[4]의 불필요한 과잉진료 등으로 보험금 지출[5]이 지속적으로 증가하는 추세라고 발표했다(MLIT, 2022). 그에 따라 금융감독원은 경상환자 과잉진료 등 자동차보험금 누수 방지를 위해 경상환자 대상으로 4주 이상의 장기치료를 받으려면 의료기관의 진단서 발급 등 자동차보험 안정화 대책을 발표하였다(FSS, 2022). 또한 Jeon(2022)은 자동차보험 경상환자 과잉진료 현황과 경제 상황에 대한 논문을 발표하였다.

본 논문에서는 Target변수로 교통사고 경상환자 진료비용 상승의 주 항목인 입원기간을 설정하였다. 자동차보험 심사 데이터와 머신러닝 알고리즘를 이용하여, 입원기간이 18일을 초과하는 환자를 예측하는 모델을 생성하고, 입원기간에 영향을 미치는 요인을 분석하고자 한다. 또한 최적의 모델 생성을 위한 Decision tree classifier, Random forest classifier, Ada boost classifier, Gradient boost classifier, XGBoost 등 5개 알고리즘별 성능을 비교·분석한다. 본 논문의 연구 결과는 경상 환자들의 과잉진료 및 사회적 보험료 비용을 줄이는 진료행태 분석에 도움이 될 것이다.

논문의 구성은 선행연구를 탐색하고 연구 데이터셋 및 전처리 과정 설명, 모델 생성 및 연구 결과 분석, 마지막 결론으로 연구 결과 의의와 향후 과제 순으로 제시할 것이다.

4)상해 12~14등급: 근육 또는 힘줄의 단순염좌, 3cm 미만의 얼굴 부위 찢김 상처 등

5)보험금 : ('16) 3.3조원→ ('21) 4.5조원, 경상보험금 : ('16) 1.9조원 → ('21) 3조원 vs 중상보험금 : ('16) 1.4조원→ ('21) 1.5조원

2. 선행연구

자동차 사고 예측 모델 개발 관련 연구는 사고 위험도, 사고 발생 패턴, 사고 감소 예측 등 사고 발생과 관련된 논문이 대부분이다. 사고 발생 후 장기입원 관련 환자의 입원기간 및 그 요인을 예측하는 해외 연구논문은 다수 있으나 국내 논문은 미흡한 편이다.

Adeleke et al.(2018)은 교통사고 피해자의 입원 기간 예측 관련 다중 선형 회귀 분석과 decision tree 기법을 분석하였다. 교통사고 피해자 450명을 대상으로 나이, 성별, 부상 유형, 병력 등 요인에 대한 데이터를 이용하여, 입원기간을 예측할 수 있는 모델을 만들었다. 분석결과 decision tree 모델의 입원 기간 예측 성능이 다중 선형 회귀 모델보다 좋은 성능을 보였다. decision tree 분석이 자동차사고 환자의 입원기간 예측, 치료 및 의료자원 할당에 활용될 수 있음을 보여주었다.

Abujaber et al.(2020)은 총 15만여 명의 환자가 포함된 47개의 연구에 대해 메타분석을 수행하였다. 7일 이상 장기입원을 하는 자동차 충돌 환자를 예측하는 요인을 분석하였다. 그 결과 나이, 여성, 부상의 심각도, 머리 부상, 척추 부상, 병원 이송 시간 등의 요인들이 장기입원과 관련 있다는 것을 제시했다. Almuheidi et al.(2021)은 자동차 사고로 입원한 환자들을 대상으로 장기입원과 관련된 요인을 조사한 34개의 연구들을 메타분석 실시했다. 장기입원을 잠재적으로 예측할 수 있는 인구통계학적, 부상 관련 및 임상적 요인에 대한 데이터를 추출하여, 장기입원에 가장 영향을 미치는 변수는 나이와 부상 심각도였으며, 머리와 척수 부상과 같은 치명적 부상도 영향이 컸다.

Byun et al.(2023)은 건강보험심사평가원 자동차보험 심사청구 한방데이터를 이용하여 40대~50대 경상환자를 대상으로 머신러닝 기반의 진료 기간을 예측하고 진료기간에 영향을 미치는 요인을 분석했다. decision tree 및 random forest 등 5개 알고리즘을 적용하여 분석한 결과 변수의 중요도는 병원의 종류, 진료지역, 나이, 성별순 이었다. 앞서 알아본 선행연구들은 머신

러닝을 이용한 자동차보험 사고 관련 진료기간 예측 연구였다. 자동차보험 사고 외 머신러닝을 이용한 암 질환 관련 선행연구도 탐색하였다. 머신러닝 기법의 입원기간 예측에 있어 자동차 사고와 암 질환 등 질병 종류에 따른 예측 모델의 정확성, 변수의 중요도 등 차이점을 분석할 필요가 있다.

Lee et al.(2023)은 유방암 건강보험 청구 데이터의 나이, 상병명, 치료횟수, 수술여부 등을 이용하여 머신러닝 기반 유방암의 생존 여부를 예측하고 생존 여부에 미치는 요인의 차이점을 분석했다. 분석 데이터 그룹을 40~50대와 60~80대로 구분하여 연령대 간의 차이점을 분석했다. 그 결과, 환자들의 생존 여부 예측 정밀도는 40~50대가 60~80대 보다 높았으며, 그 요인에 있어서도 40~50대는 치료횟수가, 60~80대는 나이의 중요도가 높았다.

교통사고 후 사고 환자의 진료기간에 대한 예측 모델에 대한 국내 논문은 초기 단계이다. 따라서 본 논문에서는 국내 자동차보험 사고 전수 데이터인 자동차보험 심사 데이터를 이용하여 사고발생 후 경상환자들의 입원기간 예측 모델 생성 및 그 요인을 분석하고자 한다.

3. 연구 데이터셋 및 전처리

3.1 연구데이터

연구대상 자료는 의료기관이 건강보험심사평가원에 자동차보험 심사 청구한 자료 중 2020년도에 심사 결정한, 상병코드 S13인 의과 입원 265,577건 요양급여비용명세서[6] 자료이다. S13은 '목부위의 관절 및 인대의 탈구, 염좌 및 긴장' 상병명으로 자동차보험 사고 다발생 1순위인 상병이다.

3.2 기계 학습을 위한 데이터 전처리

총 265,577 건의 입원환자 진료기록 중에서 입원기간이 0인 당일 입·퇴원 환자 96,459건, 결측값이 있는 31건, 날짜가 잘못 입력된 13건을 제외하고, 최종적으로 169,074건의 진료기록을 분석하였다. 진료기록 데이터는 요양기관의 종류, 소재지, 환자의 나이와 내외국인별 성별, 주상병명, 입원기간, 입원개시일, 사고발생일자, 진료비에 대한 데이터를 포함하고 있다. Target변수 설정은 입원기간이 18일을 초과하는 경우로 하였다. 이는 2021년 의료이용 현황 통계에서 입원환자 평균 재원일수가 18.5일이기 때문이다(MOHW, 2023). 통상적으로 입원 초기에 각종 검사들이 이루어지므로 입원 기간이 짧을수록 일당 진료비는 증가하며, 입원 기간이 길어질수록 총 진료비는 증가하는 경향이 나타난다. 또 진료비 자료는 퇴원시점에서 생성되므로 진료개시 시점에서 환자의 장기입원 여부를 예측하는 알고리즘에 진료비 자료를 설명변수로 설정하는 것은 적절하지 않다. 이 경우 모델의 정확도가 지나치게 과대 평가될 수 있다. 따라서 진료개시 시점에서 알 수 없으며, 평균 일당 진료비, 총진료비와 같이 입원기간 데이터 산출에 직접적인 관계가 있는 지표인 진료비 항목은 설명변수에서 모두 제외하였다.

총 169,074건의 진료 자료를 분석하였지만, 원자료가 지나치게 세분화되어 있어 한 범주에 속하는 자료가 너무 적은 경우, 이를 설명변수에 포함시키게 되면 예측 알고리즘이 수렴하지 못하고 확산되는 현상이나 과적합이 발생할 가능성이 높아진다. 따라서 원자료는 명확한 기준에 따라 범주화시켜야 한다. 요양기관의 종류는 상급병원급, 종합병원급, 병원급, 요양병원급, 정신병원급, 의원급, 보건의료원급, 한방병원급, 한의원급으로, 요양기관 소재지 정보는 서울, 부산, 인천, 대구, 광주, 대전, 울산, 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주, 기타로 범주화하였다. 외국인의 경우 관찰값이 적어 별도로 구분하지 않았고, 성별과 나이로만 구분하였다. 상병명의 경우에도 상해의 정도에 따라 추간판 외상성 파열, 탈구, 염좌로 범주화하였다. 최종적으로 설명변수로는 요양기관의 급(class), 광역 소재지, 환자 나이, 성별, 중분류 상병명을 사용하였고, 결과변수로는 입원기간의 18일 초

[6] 요양급여비용명세서 : 의료기관에서 건강보험심사평가원으로 보험 비용 청구를 위하여 환자 단위로 1주 또는 월 단위로 진료내역을 기재하여 청구하는 자료

과 여부로 설정하였다.

3.3 기계 학습을 위한 모델 세팅 및 데이터셋 설정

자료 분석은 윈도우 환경에서 visual studio code 1.80.2, python 3.8.16을 사용했다. 자료는 전체 데이터를 8대 2의 비, 무작위로 훈련 데이터(train dataset)와 실험 데이터(test dataset)로 분리하였다. 훈련 데이터는 135,259건, 실험 데이터는 33,815건이었다. 나이를 제외한 모든 데이터의 속성이 범주형이었기 때문에, 범주형 자료 분류에 적합한 ①decision tree classifier, ② random forest classifier, ③Ada boost classifier, ④gradient boost classifier, ⑤XGBoost를 기계학습 모델로 사용하였다.

예측모형의 수렴 여부를 확인하기 위해 교차검증(cross-validation)을 10회씩 실시하여, 정확도의 표준편차를 계산하였다. 예측모델의 최대 심도(max_depth)는 5단계로 제한하여 과적합이 발생하지 않도록 하였으며, 최상의 분할을 찾을 때 고려해야 할 feature의 수(max_features)는 0.6으로 설정하였고, learning rate는 0.1, n_estimators는 400, subsample은 0.7로 설정하였다. 각 기계학습 모델에서 설명변수의 중요도는 feature importance를 계산하여 평가하였다. 학습된 각 모델 알고리즘을 실험 데이터에 적용하여, 진료개시 시점이나 입원 초기에 예측 알고리즘이 장기입원환자를 얼마나 잘 예측하는지 정확도를 계산하여 평가하였다.

3.4 예측 결과에 따른 설명변수 차이 분석

인공지능 모델이 분류한 결과에 따라 예측에 사용된 설명변수에 유의한 차이가 있는지 추정하기 위하여 실험 데이터에 대한 추가 분석을 실시하였다. 추가분석을 위해서 검정력(power)을 가질 수 있도록 각 군과 설명변수에 배정되는 샘플의 수가 충분해야 한다. 따라서 요양기관의 종류는 상급병원급, 종합병원급, 병원급, 의원급으로 재분류하였다. 요양기관 소재지 정보는 수도권, 기타 광역시, 충청, 전라/제주, 경상/강원으로 더 단순화시켰다. 범주형 변수의 경우, χ^2 test를 실시하였으며 교차표에서 기대빈도가 5보다 작은 셀이 20% 이상인 경우 Fisher's exact test를 실시하여 최종 p-value를 계산하였다. 연속형 변수인 나이의 경우에는 independent samples t-test를 실시하였으며, Shapiro-Wilk test를 이용한 정규성 검증, Levene's test를 통한 군간 분산의 동질성 검증을 실시하여 t-test를 위한 가정조건을 만족하는지 여부를 확인하였고, 조건을 충족시키지 못하는 경우에 Mann-Whitney U test 결과로 최종 판단하였다. 추가 통계분석을 위해서 scipy 1.24.3 library가 사용되었으며, 유의성 판단기준은 p-value 0.05 미만이었다.

4. 연구 결과

4.1 모델별 설명변수 중요도와 예측 정확도

4.1.1 훈련 데이터의 Cross-validation

훈련 데이터에 대한 예측 모델별 교차검증 결과는 Table 1에서 확인할 수 있다. 평균 정확도는 0.913~0.915, 정확도의 표준편차는 0.001 미만으로 모델간에 큰 차이는 없었으며, 모든 모델에서 상당한 정확도로 장기입원환자를 추정하였다.

Table 1 Cross-validation results for each prediction model.

Prediction model	Accuracy mean (S.D.)
Decision tree	0.914 (<0.001)
Random forest	0.915 (<0.001)
Ada boost	0.913 (<0.001)
Gradient boost	0.914 (<0.001)
XGBoost	0.914 (<0.001)

S.D. means standard deviation.

4.1.2 예측 모델별 설명변수의 중요도

훈련 데이터에서 각 예측 모델의 정확도는 서로 비슷하였지만, 모델에 사용된 설명변수들의

중요도는 모델마다 서로 달랐다. 각 모델별 설명변수들의 중요도는 Fig. 1과 같다. 요양기관의 소재지의 경우, decision tree, random forest 모델에서는 상대적으로 덜 중요했지만, Ada boost 모델에서는 중요한 변수로 사용되었음을 확인할 수 있었다.

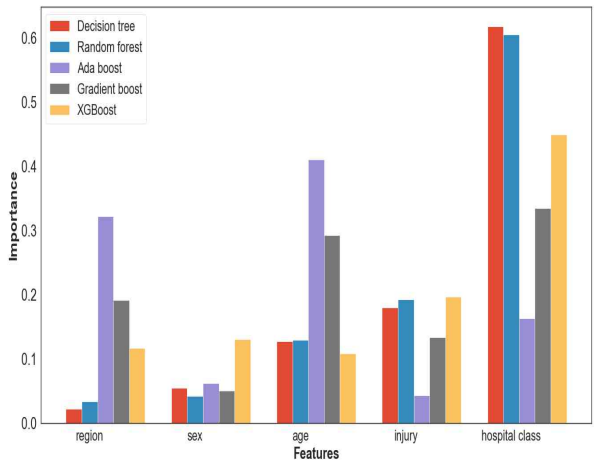

Fig. 1 Importance of each feature variable according to the models.

4.1.3 예측 모델별 성능평가

실험 데이터에서 각 예측 모델별 정확도는 Table 2에 제시하였다. 모델의 성능 평가에 있어 가장 중요하게 사용되는 정확도와 정밀도는 각 모델 사이에 큰 차이는 없었지만, random forest와 XGBoost 모델에서 가장 높은 정확도를 보였다. Recall은 0.91, F1 Score는 0.88로 5개 모델에서 모두 동일한 성능을 보였다.

4.2 모델별 예측결과에 따른 설명변수의 차이

각 모델에서 예측한 결과에 따라 사용된 변수들에 실제로 유의한 차이가 있는지 여부를 분석한 결과는 Table 3와 Table 4에 제시하였다.

4.2.1 요양기관의 소재지

실험 데이터에서 장기입원 환자와 장기입원을 하지 않은 환자 사이에 요양기관 소재지 분포에 유의한 차이가 관찰되었으며(χ^2=95.259, p<0.001),

Table 2 Performance evaluation of prediction models (Test Data Set)

Prediction model	Accuracy(%)	Precision	Recall	F1 score
Decision tree	91.448	0.89	0.91	0.88
Random forest	91.451	0.89	0.91	0.88
Ada boost	91.377	0.87	0.91	0.88
Gradient boost	91.430	0.88	0.91	0.88
XGBoost	91.451	0.89	0.91	0.88

각 예측 모델이 분류한 장기입원환자와 비장기입원환자 사이에도 지역간 유의한 차이는 확인되었다(decision tree, χ^2=23.723, p<0.001; random forest, χ^2=29.791, p<0.001; Ada boost, χ^2=12.494, p=0.014; gradient boost, χ^2=11.814, p=0.019; XGBoost, χ^2=11.073, p=0.026).

4.2.2 성별

실험 데이터에서 장기입원환자와 비장기입원환자 사이에 성별에 따라 유의한 차이가 관찰되었으나(χ^2=86.727, p<0.001), Ada boost와 gradient boost의 분류 결과 사이에는 성별에 유의한 차이가 확인되지 않았다(Ada boost, χ^2=1.143, p=0.285; gradient boost, χ^2=0.301, p=0.583). 반면에 다른 모델의 분류결과 사이에는 유의한 차이가 있었다(decision tree, χ^2=6.395, p=0.011; random forest, χ^2=4.038, p=0.044; XGBoost, χ^2=5.927, p=0.015).

4.2.3 나이

실험 데이터에서 군간에 유의한 나이 차이가 관찰되었으나(p<0.001), decision tree와 random forest의 분류결과 사이에는 유의한 차이가 없었다(decision tree, p=0.532; random forest, p=0.096). 다른 모델의 분류결과 사이에는 유의한 차이가 있었다(Ada boost, p<0.001; gradient boost, p<0.001; XGBoost, p<0.001).

4.2.4 상병명

Table 3 Differences in categorical explanatory variables according to prediction results

Variables	True value		Decision tree		Random forest	
	χ^2	p	χ^2	p	χ^2	p
Region	95.259	< 0.001	23.723	< 0.001	29.791	< 0.001
Capital area-Metro city	0.652	0.42	0.285	0.593	0.008	0.93
Capital area-Chungcheong	6.517	0.011	0.192	0.661	0.959	0.328
Capital area-Jeolla/Jeju	1.231	0.267	0.123	0.726	1.733	0.188
Capital area Gyeongsang/Gangwon	70.851	< 0.001	15.858	< 0.001	24.041	< 0.001
Metro city-Chungcheong	8.202	0.004	0.62	0.431	0.826	0.363
Metro city-Jeolla/Jeju	0.14	0.708	0.008	0.929	1.37	0.242
Metro city Gyeongsang/Gangwon	39.276	< 0.001	11.588	0.001	13.679	< 0.001
Chungcheong-Jeolla/Jeju	8.749	0.003	0.388	0.533	0.032	0.858
Chungcheong-Gyeongsang/Gangwon	60.14	< 0.001	4.105	0.043	4.672	0.031
Jeolla/Jeju-Gyeongsang/Gangwon	25.896	< 0.001	7.55	0.006	4.664	0.031
Sex	86.727	< 0.001	6.395	0.011	4.038	0.044
Injury	282.134	< 0.001	17,049.481	< 0.001	20,103.684	< 0.001
R.I.D.D[7]	0.123	0.725	0.871	0.84	2.602	0.107
R.I.D.S[8]	181.827	< 0.001	16,645.539	< 0.001	17,996.778	< 0.001
Dislocation-Sprain	102.732	< 0.001	17,798.245	< 0.001	23,422.911	< 0.001
Hospital class	786.729	< 0.001	1,234.465	< 0.001	1,597.905	< 0.001
Tertiary hospitals-General hospitals	2.264	0.132	277.149	< 0.001	382.619	< 0.001
Tertiary hospitals-Hospitals	13.446	< 0.001	931.920	< 0.001	1,057.048	< 0.001
Tertiary hospitals-Clinics	55.420	< 0.001	1,633.050	< 0.001	2,167.353	< 0.001
General hospitals-Hospitals	126.138	< 0.001	21.055	< 0.001	15.414	< 0.001
General hospitals-Clinics	773.522	< 0.001	46.126	< 0.001	48.99	< 0.001
Hospitals-Clinics	306.68	< 0.001	2.657	0.103	7.35	0.007

실험 데이터에서 장기입원환자와 비장기입원환자 사이에 상병명에 따라 유의한 차이가 관찰되었으며(χ^2=282.134, p<0.001), 각 예측 모델이 분류 결과 사이에도 모두 유의한 차이는 확인되었다(decision tree, χ^2=17,049, p<0.001; random forest, χ^2=20,104, p<0.001; Ada boost, χ^2=7,005, p<0.001; gradient boost, χ^2=11,880, p<0.001; XGBoost, χ^2=11,329, p<0.001).

4.2.5 요양기관의 종류

실험 데이터에서 장기입원환자와 비장기입원환자 사이에 요양기관의 종별에 따라 유의한 차이가 관찰되었으며(χ^2=786.729, p<0.001), 각 예측 모델 분류 결과 사이에도 모두 유의한 차이는 확인되었다(decision tree, χ^2=1,234, p<0.001; random forest, χ^2=1,598, p<0.001; Ada boost, χ^2=363, p<0.001; gradient boost, χ^2=807, p<0.001; XGBoost, χ^2=830, p<0.001).

5. 논 의

5.1 연구 결과 논의

7) R.I.D.D : Rupture of intervertebral disc - Dislocation

8) R.I.D.S : Rupture of intervertebral disc - Sprain

Table 3 Be continued.

Variables	Ada boost		Gradient boost		XGBoost	
	χ^2	p	χ^2	p	χ^2	p
Region	12.494	0.014	11.814	0.019	11.073	0.026
Capital area-Metro city	0.009	0.924	0.89	0.346	1.434	0.231
Capital area-Chungcheong	5.072	0.024	2.048	0.152	< 0.001	0.996
Capital area-Jeolla/Jeju	6.782	0.009	1.353	0.245	4.092	0.043
Capital area Gyeongsang/Gangwon	5.914	0.015	7.103	0.008	3.941	0.047
Metro city-Chungcheong	3.338	0.068	3.852	0.05	0.826	0.363
Metro city-Jeolla/Jeju	4.283	0.039	3.042	0.081	6.977	0.008
Metro city Gyeongsang/Gangwon	3.488	0.062	3.042	0.004	6.686	0.01
Chungcheong-Jeolla/Jeju	0.017	0.896	0.063	0.802	1.878	0.171
Chungcheong-Gyeongsang/Gangwon	0.028	0.867	0.351	0.553	1.624	0.203
Jeolla/Jeju-Gyeongsang/Gangwon	0.107	0.743	0.82	0.363	0.038	0.846
Sex	1.143	0.285	0.301	0.583	5.927	0.015
Injury	7,005.376	< 0.001	11,880.357	< 0.001	11,328.519	< 0.001
R.I.D.D[7]	0.878	0.839	1.143	0.84	0.703	0.421
R.I.D.S[8]	455.796	< 0.001	1,374.531	< 0.001	1,260.248	< 0.001
Dislocation-Sprain	519.040	< 0.001	1,202.714	< 0.001	1,792.653	< 0.001
Hospital class	362.607	< 0.001	807.056	< 0.001	829.928	< 0.001
Tertiary hospitals-General hospitals	77.093	< 0.001	229.327	< 0.001	193.615	< 0.001
Tertiary hospitals-Hospitals	265.028	< 0.001	564.324	< 0.001	588.633	< 0.001
Tertiary hospitals-Clinics	538.954	< 0.001	1,175.952	< 0.001	1,389.633	< 0.001
General hospitals-Hospitals	30.556	< 0.001	12.716	< 0.001	20.339	< 0.001
General hospitals-Clinics	79.573	< 0.001	45.938	< 0.001	66.383	< 0.001
Hospitals-Clinics	7.884	0.005	8.739	0.003	10.457	0.001

본 연구는 2020년도 자동차보험 심사 청구자료를 이용하였다. 자동차보험 사고 다발생 1위인 S13 상병명(목부위의 관절 및 인대의 탈구, 염좌 및 긴장)코드로 의과(한방 제외)에 입원한 환자 169,074건을 분석하였다. 기계학습을 위한 모델 셋팅은 8대2 비율로 훈련 데이터와 실험 데이터로 구분하여 실험했다. 모델 알고리즘은 decision tree, random forest, Ada boost, gradient boost, XGBoost를 사용하여 모델을 구축했다. 예측 모델의 평가는 훈련 데이터와 실험 데이터의 장기입원환자 예측의 정확도를 계산하여 평가하였다. 모델의 설명변수 중요도는 실험 데이터의 feature importance를 적용하였다. 연구 결과, 실험 데이터에 대한 5개 예측 모델의 정확도는 91.377~91.451이며 각 모델 사이에 큰 차이점은 없었으나 random forest와 XGBoost 모델이 91.451로 가장 높았다. 모델별 설명변수의 중요도는 다음과 같다. 요양기관의 소재지(χ^2=95.259, p<0.001), 상병명(χ^2=282.134, p<0.001), 요양기관 종류(χ^2=786.729, p<0.001)는 장기입원을 한 환자와 장기입원을 하지 않은 군 사이에 모든 예측 모델에서 유의한 차이를 나타냈다. 하지만 성별, 나이는 비교군 사이에 유의한 차이가 관찰되었으나 모델별 분류 결과 사이에는 유의한 차이는 확인되지 않았다.

본 논문의 연구 방법과 유사한 선행 연구인

Table 4 Age difference according to prediction results of each model

	True value	Decision tree	Random forest	Ada boost	Gradient boost	XGBoost
LTH, mean (S.D.)	50.60 (15.05)	47.00 (12.42)	51.23 (11.99)	55.15 (18.88)	56.79 (17.81)	57.01 (17.39)
Non-LTH, mean (S.D.)	47.06 (15.62)	47.36 (15.61)	47.36 (15.61)	47.34 (15.59)	47.34 (15.59)	47.34 (15.60)
Independent samples t-test						
t	-11.701	0.175	-2.000	-5.194	-5.611	-5.291
p	< 0.001	0.861	0.046	< 0.001	< 0.001	< 0.001
Levene's test						
F	18.980	7.835	8.816	7.973	4.074	1.731
p	< 0.001	0.005	0.003	0.005	0.044	0.188
Mann-Whitney U test						
p	< 0.001	0.532	0.096	< 0.001	< 0.001	< 0.001

LTH means long-term hospitalization; S.D., standard deviation.

Byun et al.(2023)의 연구 결과는 다음과 같다. 입원 기간 11일 이상 예측률에 있어 decision tree 등 5개 모델의 평균 정밀도는 0.78이었다. 설명변수 중요도에 있어 입원 기간 예측률에 영향을 미치는 요인 분석 결과, 병원구분_한방병원(0.839), 진료지역_광주·전남·전북(0.093), 병원구분_한의원(0.021), 진료지역_부산·경남·울산(0.008) 순이었다.

본 논문과 Byun et al.(2023)의 예측 모델 평가에 있어서 정확도(약 91%)와 정밀도(약 78%)로 평가 측도의 차이점은 있었으나 본 논문의 정확도가 수치상 높았다. 설명변수의 중요도는 요양기관 소재지(진료지역)와 요양기관의 종류(병원구분)는 각 연구에서 중요 요인으로 유사하게 나타났다.

5.2 연구 한계

전체 입원환자 데이터 중에서 입원기간이 18일을 초과하는 장기입원환자의 비율은 9.39%였다. 이렇게 장기입원과 비장기입원 환자군의 샘플 수에서 큰 차이를 보이는 경우, 양성 예측도나 음성 예측도와 같은 평가지표들이 과대 또는 과소 평가되는 경향이 나타난다(Park et al., 2020). 따라서 군에 속하는 샘플의 수를 인위적으로 맞추어 분석하여 정확한 예측력을 구하는 방법이 사용될 수도 있다. 그러나 이 경우 실제 임상환경에서 모델을 적용하면, 기대되었던 만큼 정확도가 나오지 않는다. 따라서 실제 데이터의 발생빈도에서 차이가 있는 경우에는 검증도 유사한 발생빈도 조건에서 검증이 이루어져야 한다. 본 연구에서는 건강보험심사평가원 자료를 무작위로 훈련 데이터와 실험 데이터를 나누어 모델을 개발한 환경과 실험 환경에 사용된 데이터가 최대한 동일한 조건이 되도록 설정하였으며, 모델의 정확도가 과대 평가되지 않도록 하였다.

설명변수 구성에 있어 환자 진료 정보 중 가장 기초적인 데이터인 요양기관의 종류와 소재지, 환자의 나이와 성별, 상병명만을 활용하여 입원기간을 추정하였다. 실제로 입원기간은 환자의 기저질환의 종류와 유무, 환자의 경제적인 상황, 요양기관의 접근성 등과 같은 다양한 외생변수들이 작용하여 결정될 것이다. 이러한 외생변수들을 예측 모델에 추가한다면 더 정확한 예측도 가능할 것이다.

6. 결 론

본 연구는 자동차 사고 환자를 대상으로 공공데이터를 기계학습법을 통해 분석한 초기 연구로서, 예측 결과는 사용된 기계학습 모델의 종류와 상관없이 모두 상당한 정확도를 나타냈다. 그러나 예측에 사용된 설명변수들의 중요도는 모델마다 상당한 차이가 있었다. 이러한 차이로 인해 decision tree와 random forest 모델의 경우 나이에 따른 입원기간의 차이를 반영하지 못했고, Ada boost와 gradient boost 모델의 경우 성별에 따른 입원기간의 차이를 반영하지 못했다. XGBoost 모델의 경우에는 모든 설명변수에 따른 입원기간의 차이를 예측하였으며, 예측 정확도도 다른 모델보다 높아 진료 초기에 장기입원환자를 비교적 정확하게 추정하는 모델이었다.

자동차보험금 관리기관인 국토교통부, 금융감독원 등은 효과적인 보험금 지출을 위해 교통사고 경상환자의 장기 진료행태 개선을 위해 노력하고 있다(FSS, 2022; Jeon, 2022; MLIT, 2022). 본 논문은 자동차사고 경상환자 데이터의 방대함으로 외래를 제외한 입원 건을 대상으로 장기입원을 예측하는 모델 생성 및 그에 미치는 요인을 분석했다. 그 결과 의료기관은 입원단계에서, 자동차보험 관리기관과 심사기관은 사후관리 단계에서 적정 입원기간에서 벗어난 환자 사례를 감지하는데 도움될 것으로 본다. 본 논문이 자동차사고 경상환자의 장기입원 등 과잉진료를 방지하기 위한 문제 해결의 초기 연구논문으로 자리매김하기를 기대한다. 향후에는 보험금 지출이 더 많은 외래건, 검사내용 및 세부 진료내용 등 더 많은 진료행태 연구변수들을 활용한 연구가 수행되기를 기대한다.

References

Abujaber, M. A. F., Al-Majali, O., Azab, M., Ababneh, B., Alqasrawi, O., Abusamak, M., Al-Hadidi, D. and Al-Mousa, D. (2020). Predictors of prolonged hospital stay following a motor vehicle crash: A systematic review and meta-analysis, *Injury Epidemiology*, 7(1), 50

Adeleke, I. F., Adebiyi, M. O. and Adekunle, O. B. (2018). Predicting the length of hospital stay of road traffic accident victims: A comparative study between regression analysis and decision tree techniques, *Journal of Public Health,* 26(4), 375-382.

Almuheidi, S., Alaklabi, A., Miroshnichenko, A. and Almutairi, A., (2021). Prediction of hospital stay in acute traumatic brain injury: A machine learning approach, *Journal of Head Trauma Rehabilitation*, 36(1), E31-E37.

Byun, K. K., Lee, D. G. and Lee, H. D. (2023). Machine Learning-Based Prediction Technology for Medical Treatment Period of Automobile Insurance Accident Patients, *Convergence security journal,* Vol.23 No.1, 89-95

FSS. (2022). Reorganization of compensation process in line with the implementation of measures for auto insurance miner patients, Press Release(Reported on Dec. 29, 2022.)

HIRA. (2022). Car Insurance Medical Expenses Statistics

Jeon, Y. S. (2022). Changes in the Economic Environment and Over-Treatment of Minor Patients in Auto Insurance, *KIRI Report*, Vol. 542, 8-13.

KoROAD. (2023). Traffic Accident Analysis System.

Lee, D. G., Byun, K. K., Lee, H. D. and Shin, S. H. (2023). The Prediction of Survival of Breast Cancer Patients Based on Machine Learning Using Health Insurance Claim Data, *Korea Industrial Information Systems Society Journal,* 28(2), 1-9.

MLIT. (2022). Prevention of leakage of car insurance money, such as excessive medical treatment for minor patients, Press Release, (Reported on Aug. 7th, 2022)

MOHW. HIRA. (2023). Healthcare utilization statistics for 2021

Kang, S. O., Kim, S. H. and Ryu, M. H. (2022). Analysis of Hypertension Risk Factors by Life Cycle Based on Machine Learning, *Korea Industrial Information Systems Society Journal*, 27(5), 73-82.

Park, D. A., Hwang, J. S., Lee, S. H., Sul, A. R., Choi, W. J., Oh, S. H., Lee, J. Y., Lee, Y. K., Lee, D. H. and Choi, S. G. (2014). Systematic review of diagnostic tests. National Evidence-based Healthcare Collaborating Agency.

Yun, S. O., Jung, J. G., Wo, H. G. and Kim, J. E. (2022). Breast Cancer Survival Prediction: Model Comparison and Effect of Genetic Features, *Database Research,* 38(1), 3-15.

이 덕 규 (DoegGyu Lee)

- 정회원
- 가톨릭대학교 의료경영대학원 의료경영학과 석사
- (전)건강보험심사평가원 실장
- (현재)숭실대학교 IT정책경영학과 박사과정
- 관심분야: 보건의료 관련 IT정책 및 정보화분야

남 동 현 (DongHyun Nam)

- 경희대학교 한의학 박사
- 건강보험심사평가원 진료심사평가위원회 위원
- (현재) 식품의약품안전처 의료기기위원회 위원
- (현재) 식품의약품안전처 국가표준 한의약 전문위원회 위원
- (현재) 상지대학교 한의과대학 한의예과 교수
- 관심분야 : 인공지능, 의공학, 진단용 의료기기, 디지털 헬스케어

허 성 필 (Sung-Phil Heo)

- 종신회원
- 도호쿠대학교 정보통신공학 박사
- KT 연구소 (수석연구원, 부장)
- 금오공과대학교 ICT융합연구센터 교수
- 경운대학교 무인기공학과 부교수, 학과장, 공용장비지원센터 센터장
- (현재) 강릉원주대학교 교수, 산학융합지구사업단 기획전략센터 센터장
- 관심분야 : 디지털 헬스케어, 사물인터넷(IoT), 인공지능, 멀티미디어검색, 차세대 무선통신기술