KAICTS Conference of Korea Artificial-Intelligence Convergence Technology Society

합성 데이터를 활용한 머신러닝 모델의 의료 보험료 예측 최적화

지정원[1], 김준화[2]
[1]건양대학교 의료인공지능학과
[2]건양대학교 인공지능학과

Ⅰ. 서론

최근 몇 년간 의료 수술비용의 급증과 함께 의료 보험료도 함께 상승하고 있어, 개인과 가정의 경제적 부담이 커지고 있다 [1]. 이러한 상황 속에서 의료 보험료의 정확하고 신뢰할 수 있는 예측은 필수적이다. 머신러닝 기법이 발전함에 따라 다양한 변수와 패턴을 고려한 예측이 가능해졌으며 [2], 이를 활용한 의료 보험료 예측은 더욱 효과적일 것으로 기대된다.

본 논문은 머신러닝 기법을 통해 의료 보험료 예측의 정확성을 높이고자 하는 목표를 가지고 있으며, 이를 위해 다양한 알고리즘을 적용하여 예측 모델을 최적화했다. 특히, 데이터 상관관계 분석을 통해 상관관계가 높은 특성(feature)들을 합성하여 새로운 열을 생성함으로써, 실험 결과가 보다 정교하고 신뢰할 수 있는 의료 보험료 예측 결과를 도출하는 데 기여할 것으로 기대된다.

Ⅱ. 본론

2.1. 데이터 수집

의료 보험료 예측에 필요한 데이터세트는 kaggle의 Medical Cost Personal Datasets [3]를 사용하였다. 이 데이터세트는 10개의 Feature인 나이(Age), 당뇨병(Diabete), 혈압 문제(Blood Pressure Problems), 장기 이식 여부(Any Transplants), 만성 질환 여부(Any Chronic Diseases), 키(Height), 몸무게(Weight), 알레르기 여부(Known Allergies), 암 가족력(History of Cancer in Family), 주요 수술 횟수(Number of Major Surgeries)와 1개의 label인 보험료(Premium Price)로 구성되어 있다.

데이터 행 개수는 총 986개로, 학습 데이터 개수는 788개, 테스트 데이터 개수는 198개로 나누었다.

2.2. 모델 학습

첫 번째 실험에서는 여러 모델을 사용해보며 전처리되지 않은 기본 데이터세트로 실험을 진행했다.

성능 평가 지표는 R-squared [4]를 사용했고, 이는 회귀 모델에서 독립 변수가 종속 변수를 얼마나 잘 설명해주는지 보여준다. R-squared는 수식 (1)과 같이 나타내며,

$$R^2 score = \frac{SSE}{SST} = 1 - \frac{SSR}{SST} \qquad (1)$$

SST(Total Sum of Squares)는 제곱의 총합, SSE(Explained Sum of Squares)는 회귀 제곱합, SSR(Residual Sum of Squares)은 잔차 제곱합을 의미한다.

2.3. 합성 데이터

두 번째 실험에서는 첫 번째 실험에서 도출된 최고 모델의 성능을 높이기 위해 합성 데이터를 생성하였다.

이를 생성하기 위해 그림 1, 2와 같이 Premium Price와 모든 feature 간의 상관관계 분석을 통하여 어떤 특성의 상관관계 계수가 높은지 분석했고, 그중 값이 높은 장기 이식 여부(Any Transplants), 주요 수술 횟수(Number of Major surgeries), 만성 질환 여부(Any Chronic Diseases) 세 개를 합성하여 두 개의 새로운 열을 추가했다.

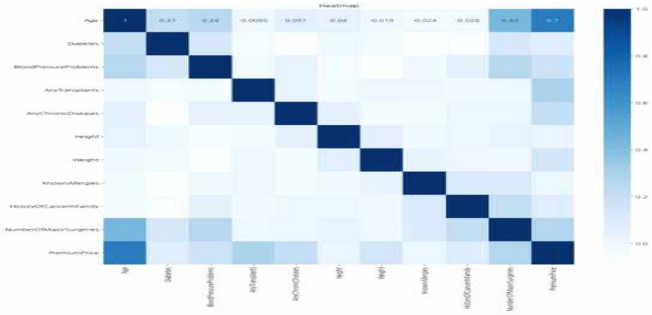
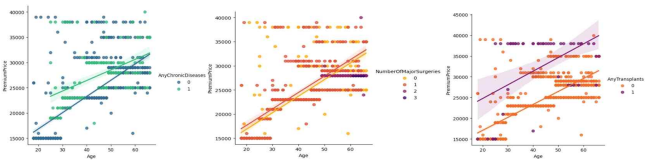


그림 1. feature 간의 상관계수를 나타낸 Heatmap



(a) (b) (c)

그림 2. (a) 만성 질환 여부와 보험료와의 상관관계(0.208), (b) 주요 수술 횟수와 보험료와의 관계(0.262), (c) 장기 이식 여부와 보험료와의 상관관계(0.289). x축은 나이, y축은 보험료.

2.4. 하이퍼파라미터 최적화

세 번째 실험에서는 두 번째 실험 결과에서의 성능을 향상시키기 위해 하이퍼파라미터 최적화 기법인 Optuna를 활용하여 최적의 하이퍼파라미터를 도출하였다. Optuna는 효율적인 하이퍼파라미터 탐색을 지원하는 자동화된 최적화 프레임워크로, 적은 수의 실험으로도 최적의 하이퍼파라미터 조합을 찾을 수 있는 장점이 있다.

Ⅲ. 실험 결과

표 1. 머신러닝 모델 결과

Model	RMSE	MAE	R^2
Decision Tree	2,271	1,368	0.878
Random Forest	2,047	991	0.901
Linear Regression	3,495	2,586	0.713
GBM	2,278	1,121	0.878
LightGBM	2,255	1,387	0.880
AdaBoost	2,924	2,047	0.799

표 2. 합성데이터 추가 모델 결과

Model	RMSE	MAE	R^2
GBM [5]	2,249	1,201	0.881
Random Forest [6]	2,075	944	0.898

표 3. 합성데이터 추가 모델 하이퍼파라미터

Model	Hyper parameter
GBM	• max_depth(10) • n_estimators(50) • min_samples_leaf(10) • learning_rate(0.1)
Random Forest	• max_depth(30) • n_estimators(70) • min_samples_leaf(2) • min_samples_split(2)

표 4. Optuna 활용 모델 결과

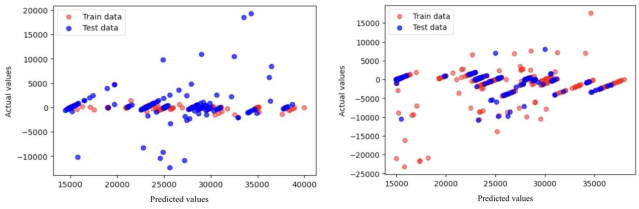
Model	RMSE	MAE	R^2
GBM	2,148	1,192	0.891
Random Forest	2,041	932	0.902

표 5. Optuna를 적용한 하이퍼파라미터 최적화

Model	Hyper parameter
GBM	• max_depth(6) • n_estimators(201) • min_samples_leaf(5) • learning_rate(0.02)
Random Forest	• max_depth(13) • n_estimators(365) • min_samples_leaf(1) • min_samples_split(10)

표 1은 기본 데이터세트로 여러 모델을 실험한 첫 번째 결과를 제시하며, 표 2, 표 3은 합성 데이터가 추가된 두 번째 실험 결과와 하이퍼파라미터를 나타낸다. 표 4, 표 5는 Optuna를 활용하여 수행한 세 번째 실험 결과와 최적의 하이퍼파라미터를 나타낸다.

최종적으로 세 가지 실험을 비교한 결과, 마지막 실험에서 GBM, Random Forest 모델의 성능이 가장 높게 나타났으며, 이를 가장 낮은 성능과 비교했을 때 R-squared 값이 각각 0.013, 0.004 증가한 것을 확인할 수 있다. 그림 3은 세 번째 실험에서 도출된 예측값의 오차 정도를 보여주는 그래프이고, 이를 통해 모델의 예측 정확성을 시각적으로 평가할 수 있다.



(a) (b)

그림 3. (a) GBM 모델에서 예측값과 실제값의 차이, (b) Random Forest 모델에서 예측값과 실제값의 차이

Ⅳ. 결론

본 논문에서는 머신러닝 기법을 활용하여 의료 보험료 예측의 정확성을 높이기 위한 다양한 알고리즘의 적용과 비교 분석을 수행하였다.

실험 결과, 합성 데이터를 추가한 Random Forest 모델에서 가장 우수한 성능을 보였으며, 위의 기법을 통해 데이터의 다양성을 확보하여 모델의 일반화 능력을 향상시켰고, 다양한 상황에서 안정적인 예측이 가능하도록 하였다.

향후 연구에서는 보다 다양한 변수와 데이터세트를 포함하여 예측 모델 성능을 추가적으로 개선할 수 있는 가능성이 있다.

ACKNOWLEDGMENTS

"본 연구는 2024년 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학사업 지원을 받아 수행되었음"(2024-0-00047)

REFERENCES

[1] Y. Jung and S. I. Huh, "Changes in the level and the composition of health expenditures by income levels", vol. 18, no. 4, pp. 21-39, 2012.

[2] M. B. Savadatti and M. Dhivya, "An Overview of Predictive Analysis based on Machine learning Techniques", IEEE Conf., 2022.

[3] https://www.kaggle.com/datasets/mirichoi0218/insurance

[4] https://scikit-learn.org/1.5/modules/generated/sklearn.metrics.r2_re

[5] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine", vol. 29, no. 5, pp. 1189-1232, 2008.

[6] Leo Breiman, "Random Forests", volume. 45, pp. 5-32, 2001.