



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

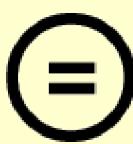
다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원 저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리와 책임은 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#)



미디어학 석사학위 논문

머신러닝 모델 해석 시스템
: 입력 변수 변화가 모델 예측에
미치는 영향을 탐색하는 시각적 분석
방법을 중심으로

아주대학교 대학원

라이프미디어 협동과정

박찬희

머신러닝 모델 해석 시스템
: 입력 변수 변화가 모델 예측에
미치는 영향을 탐색하는 시각적 분석
방법을 중심으로

지도교수 이 경 원

이 논문을 미디어학 석사학위 논문으로 제출함.

2022년 2월

아주대학교 대학원

라이프미디어 협동과정

박찬희

박찬희의 미디어학 석사학위 논문을 인준함.

심사위원장

이경원

Jae-yeon

심사위원

김효동

Kim Hyo-dong

심사위원

박홍석

Sang-huk Park

아주대학교 대학원

2022년 2월

국문 초록

해석가능한 머신러닝(interpretable machine learning)은 머신러닝 시스템의 행동과 예측을 사람이 이해할 수 있도록 돋는 기술을 말한다. 머신러닝 해석 분야에서 입력 데이터와 모델 결정 사이의 관계를 잘 설명할 수 있을 때, 과학적 이해, 안정성 파악, 신뢰 확보와 같은 이점을 얻을 수 있다. 머신러닝이 적용되는 분야가 다양해지고, 모델 해석에 대한 요구가 늘어나면서 머신러닝 모델 해석의 주체가 머신러닝 전문가에서 다양한 계층으로 확대되고 있다. 이에 따라, 모델 해석을 위해 데이터를 효과적으로 분석할 수 있는 도구 개발의 필요성이 높아졌다. 대화형 그래픽 방식을 사용하는 시각적 분석 기법은 비전문가를 비롯한 다양한 계층의 사용자가 쉽게 데이터를 분석하도록 돋는다. 따라서, 시각적 분석 기법의 도입은 머신러닝 해석 작업에서 큰 잠재력을 가지고 있다. 본 연구는 사용자가 쉽고 명확하게 머신러닝 모델을 해석할 수 있도록 지원하기 위하여 머신러닝 모델이 입력 데이터로부터 출력 결과를 어떻게 연결 짓는지에 대한 관계성을 해석하는 시각 분석 시스템을 제안한다. 기존 머신러닝 해석 연구의 한계를 극복하기 위하여 두 가지 기법이 도입되었다. 첫째, 데이터 셋의 입력 데이터 값을 합리적인 범위 내에서 변화시키고 이에 따른 모델 예측 결과 변화 추이를 추적했다. 둘째, 분석 시스템의 사용성을 높이기 위하여 유저 인터페이스에 평행 좌표 그래프 및 산점도와 같은 데이터 시각화 기법과 함께 필터, 그룹, 정렬 등의 인터랙션을 도입했다. 본 연구가 제안한 시각 분석 시스템은 머신러닝 수행 결과를 입력 변수, 목표 변수, 예측 값에 따라 필터링하고 그룹 지어 해석할 수 있는 반복적인 조정 절차를 통해 효과적으로 머신러닝 모델을 해석할 수 있는 접근 방식을 취한다. 이 시스템은 사용자가 이 분석 시스템을 사용하여 머신러닝 모델의 복잡한 동작에 대한 통찰을 얻고, 입력 변수와 목표 변수 및 모델 예측에 대한 과학적 이해를 확보하고, 모델의 안정성과 신뢰성을

파악하는데 도움을 제공한다. 유스 케이스 분석을 통해 제안된 시스템이 머신러닝 모델 해석에 도움을 줄 수 있는지 설명했다. 나아가 사용자 심층 인터뷰를 통해 제안된 시각화 및 인터랙션 기법의 제공 여부가 시스템 사용성 및 모델 해석 용이성에 미치는 영향을 평가했다. 본 연구에서 제시된 시각 분석 시스템을 통해 머신러닝 모델 해석 현장에서 발생하는 문제를 보다 쉽고 빠르게 분석할 수 있었을 뿐만 아니라 이들 문제에 대한 고도의 인사이트를 도출할 수 있음을 확인했다.

주제어: 해석가능한 머신러닝, 시각 분석 시스템, 데이터 시각화, 데이터 시각 분석, 머신러닝 모델 분석

본문 목차

제 1 장 서론	1
제 2 장 문헌 고찰	4
제 1 절 머신러닝 해석 방식의 분류	4
제 2 절 입력 변수 값 변형을 활용한 모델 해석	12
제 3 장 연구의 목적과 방법	17
제 4 장 시각 분석 시스템 설계 및 개발	21
제 1 절 평행 좌표 그래프 및 산점도	22
제 2 절 인스턴스 변형과 테이블 시각화	25
제 3 절 개별 인스턴스 확인	28
제 5 장 유스 케이스 분석	29
제 1 절 유스 케이스 지침 설계	29
제 2 절 유스 케이스 분석 수행	31
제 6 장 사용자 심층 인터뷰	37
제 1 절 사용자 심층 인터뷰 설계	37
제 2 절 사용자 심층 인터뷰 수행 및 결과	41
제 7 장 결론	45
참고문헌	47
ABSTRACT	53

그림 목차

그림 1. 의사결정나무의 예측 과정을 시각화한 BaobabView	10
그림 2. 인공 신경망 모델의 의미론적 해석을 지원하는 Testing with Concept Activation Vectors (TCAV) 방법론	11
그림 3. CNN 모델의 구조적 특성을 시각화하여 모델의 작동 원리 및 예측 결과를 해석한 CNN explainer	12
그림 4. 모델 예측의 정오를 클래스별로 시각화하여 모델이 혼동하는 클래스를 탐색하는 시각화 Squares	12
그림 5. 구역 구분을 통해 모델(B, C), 클래스(D), 인스턴스(E) 단위의 해석을 각각 제공하여 전역적 분석과 국소적 분석을 모두 가능하도록 지원하는 인터랙티브 시각 분석 시스템 사례	14
그림 6. 하이퍼파라미터 변화가 모델 성능에 주는 영향을 중심으로 모델을 시각화한 HyperTendril	15
그림 7. 인공 신경망 모델을 근사하는 단순한 구조의 규칙 기반 모델을 생성하여 이를 시각화하는 RuleMatrix	16
그림 8. 입력 값의 변화에 따른 예측 값 변화 추세를 보여주는 PD Plots	18
그림 9. 입력 값의 변화에 따른 개별 인스턴스의 예측 변화를 보여주는 ICE Plots	20
그림 10. 인스턴스를 그룹화하여 입력 변수와 모델 예측 사이에 상관관계가 존재하는지 파악하도록 돋는 DECE	21
그림 11. 연구 결과인 시각 분석 시스템을 통해 머신러닝 모델을 분석하는 과정	26
그림 12. 본 연구가 제안하는 시각 분석 시스템의 인터페이스	27
그림 13. 제안된 시스템의 평행 좌표 그래프와 산점도	28
그림 14. 필터가 적용된 평행 좌표 그래프와 스캐터플롯	31
그림 15. 다중 필터가 적용된 평행 좌표 그래프와 스캐터플롯	31
그림 16. 제안된 시스템의 테이블 시각화	32
그림 17. BMI 가 높아진 인스턴스 그룹을 표시한 테이블의 행	33

그림 18. 테이블 시각화에서 행을 클릭하면 표시되는 해당 그룹에 속한 인스턴스 목록을 보여주는 팝업 창	34
그림 19. 유스 케이스에 해당하는 데이터를 시각화한 모습	36
그림 20. 타깃 변수의 모델 예측 값이 높은 인스턴스를 필터링한 모습	37
그림 21. 타깃 변수의 모델 예측 값이 실제 값보다 큰 인스턴스를 필터링한 모습	38
그림 22. 흡연 여부가 거짓에서 참으로 변경된 그룹	39
그림 23. 흡연 여부가 참에서 거짓으로 변경된 그룹	39
그림 24. BMI 가 높은 그룹을 평행 좌표 그래프에서 필터링하고, 이들의 흡연 여부가 음에서 양으로 바꾼 그룹을 테이블 맨 위로 정렬시킨 모습	40
그림 25. BMI 가 높은 인스턴스의 흡연 여부를 음에서 양으로 바꾼 그룹의 상세 보기	41
그림 26. 제안된 시스템에 Pima Indian Diabetes 데이터 셋과 이를 학습한 로지스틱 회귀 모델이 적용된 모습	46

제 1 장 서론

머신러닝 모델은 높은 성능에도 불구하고 특정 예측을 내렸는지에 대한 명확한 설명을 제공하지 않는다는 측면에서 블랙박스라고 불린다. 해석가능한 머신러닝(Interpretable Machine Learning)분야의 목표는 머신러닝 시스템의 행동과 예측을 사람이 이해할 수 있도록 만드는 것이다¹. 머신러닝 연구에서 해석가능성의 의미를 명확히 이해하기 위하여 설명가능성(Explainability)과의 비교가 도움이 된다². 우선, 해석가능성은 모델이 입력 데이터와 예측 값 사이의 인과관계를 어떻게 연결 짓는지를 보여주는 능력이다. 비슷한 입력에서 출력 또한 비슷하다면 모델이 해석가능하다고 여긴다. 해석가능성이 높을 때, 모델의 사용자는 모델을 신뢰할 수 있다. 반면, 설명가능성은 모델 파라미터의 영향과 모델이 수행하는 프로세스를 사람이 이해가능하도록 만드는 능력이다. 모델이 결정을 내리는 데 어떤 요소가 영향을 주었는지 쉽게 알 수 있을 때, 그 모델이 설명가능하다고 여긴다. 모델의 설명가능성이 높으면 연구자나 개발자는 모델을 디버깅하기 쉽다. 설명가능성이 개발자가 모델의 성능을 개선하는 것과 연관되는 반면, 해석가능성은 최종 사용자가 모델을 신뢰하게 만드는데 관련이 깊다.

몇몇 경우에는 머신러닝 시스템이 내린 결정에 대한 이유가 중요하지 않을 수도 있다. 예를 들어, 영화 추천 시스템과 같이 실수가 심각한 결과를 초래하지 않거나, 광학 문자

¹ Molnar, Christoph. Interpretable machine learning. Lulu.com, 2020.

² Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018.

인식과 같이 이미 광범위하게 연구되고 평가된 분야가 그러하다^{3,4}. 이러한 분야에서 모델을 평가하는 데 주로 사용되는 지표는 예측 정확도를 비롯한 성능이다. 하지만, 머신러닝 시스템이 내리는 결정이 사회에 큰 영향을 끼치는 경우가 늘어나면서 시스템이 결정을 내리는 원리를 분석해야 할 필요성이 대두되고 있다. 예를 들어, 자율주행 자동차를 위한 보행자 인식 시스템이나, 치료 방향 결정을 위한 질병 검진 시스템이 그렇다^{5,6}.

기술 발전에 따라 인공지능이 인간 사회에 나쁜 영향을 미칠 수 있다는 의견이 있다. 예를 들어 Kurzweil은 인공지능을 개발하고 이용하는 주체가 인공지능이 인간의 윤리와 가치를 고려하게 만들겠다는 사회적 신뢰를 구축하는 것이 어렵다고 지적했다⁷. 모델의 안정성을 파악하거나 신뢰하기 어렵기 때문에 머신러닝 모델이 도출한 결과를 무조건 받아들이는 것이 인간의 윤리와 공정성을 침해할 수 있다는 의견이다. 반면, 인간과 인공지능이 공존할 수 있다는 긍정적인 시각도 있다. 예를 들어, Thiel은 인공지능이 인간을 대체하는 것이 아니라 보완한다고 말했다⁸. 데이터를 바탕으로 인간이 필요로

³ Vilakone, P., Park, D. S., Xinchang, K., & Hao, F. An efficient movie recommendation algorithm based on improved k-clique. *Human-centric Computing and Information Sciences* 8.1: 1–15, 2018.

⁴ Tanveer, Muhammad Suhaib, Muhammad Umar Karim Khan, and Chong-Min Kyung. Fine-tuning darts for image classification. 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.

⁵ Feifel, Patrick, Frank Bonarens, and Frank Koster. Reevaluating the Safety Impact of Inherent Interpretability on Deep Neural Networks for Pedestrian Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

⁶ Das, D., Ito, J., Kadokawa, T., & Tsuda, K. An interpretable machine learning model for diagnosis of Alzheimer's disease. *PeerJ* 7 (2019): e6543, 2019.

⁷ Kurzweil, R. *The singularity is near: When humans transcend biology*; Penguin, 2005.

⁸ Thiel, P.A.; Masters, B. *Zero to one: Notes on startups, or how to build the future*; Currency, 2014.

하는 의미 있는 결과를 만들어내기 위해서는 인공지능이 도출한 결과를 종합적인 인간의 사고로 검토해야 한다. 본 연구는 인간과 인공지능 모델이 협력하는 방식의 한 예시를 제안한다. 인공지능은 머신러닝 모델을 통해 데이터 셋 내의 입력 변수와 목적변수 사이의 관계성을 계산한다. 본 연구가 제안하는 시스템은 머신러닝 모델이 학습한 내용을 인간이 해석할 수 있는 시각화 및 인터랙션을 제공한다. 머신러닝 모델의 사용자가 모델의 결과를 그대로 따르는 것이 아니라 비판적으로 수용하는 것이다.

머신러닝 분야에서, 입력 데이터와 모델 결정의 사이의 관계를 잘 설명할 수 있을 때, 다음과 같은 이점을 얻게 된다⁹. (1) 과학적 이해: 머신러닝 모델이 학습한 바를 설명하여 입력 데이터와 출력 결과 사이의 관계에 대한 지식을 얻을 수 있다. (2) 안전성: 작은 입력 값 변화가 큰 예측 오류로 이어지지 않게 한다. (3) 신뢰: 블랙박스에 비해 결정을 설명하는 시스템을 신뢰하기 더 쉽다. (4) 공정성: 인종이나 성별을 비롯한 편견이 모델에 학습되어 있는지 확인할 수 있다.

본 연구는 머신러닝 모델의 행동과 예측을 모델 사용자가 이해할 수 있도록 돋기 위한 시각 분석 시스템을 제안한다. 특히, 인공지능이 도출한 결과를 인간의 사고로 검토할 수 있도록 만들기 위해, 머신러닝 시스템이 도출한 결과와 머신러닝 모델 해석 기법을 시각적으로 표현하였다. 2 장에서는 머신러닝 모델 해석 분야의 기존 연구와 그 한계를 살펴본다. 3 장에서는 머신러닝 해석의 공통된 목표를 달성하고 기존 연구의 한계를 극복하는 방향으로 시각화 시스템의 설계 목표 사항을 정립한다. 4 장에서는 시스템의 설계 목표를 충족시키기 위한 분석 프레임워크 및 시스템을 설계하고 안내한다. 5 장에서는 ‘건강 보험료 지출 예측 모델’의 유스 케이스를 통해 시스템의 활용 방법을

⁹ Doshi-Velez, Finale, and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.

소개하고, 시스템이 설계 목표 사항을 달성했는지 여부를 확인한다. 6 장에서는 심층 인터뷰를 통해 사용자가 본 연구가 제안하는 시스템을 통해 ‘당뇨병 예측 모델’의 분석을 수행하는 과정을 확인하고, 시스템을 사용성, 기여점 및 한계점을 평가한다. 마지막으로 7 장에서는 본 논문의 주요 결과를 요약하고 향후 연구 방향을 소개한다.

제 2 장 문헌 고찰

제 1 절 머신러닝 해석 방식의 분류

머신러닝 모델을 해석하기 위한 방법론 및 시각 분석 시스템 연구가 수행되어 왔다. 본 연구에서는 이 연구들을 해석 범위 및 목표 사용자에 맞춰 분류한다.

머신러닝 분야에는, 종종 정확도를 비롯한 성능과 해석가능성 사이의 반비례 관계가 보이곤 한다. 구조가 단순한 모델은 그 자체적으로 이미 해석력을 확보하고 있다. 의사결정나무나 선형회귀와 같은 단순한 모델은 시각화하고 해석하기 쉽다^{10 11}. 하지만 이러한 모델은 복잡한 문제에서 정확도가 떨어진다는 문제를 갖는다. 높은 성능으로 주목받고 있는 심층 신경망 모델은 높은 성능을 지니지만 모델 자체가 설명력을 가지지

¹⁰ Van Den Elzen, S., & Van Wijk, J. J., Van Den Elzen, Stef, and Jarke J. Van Wijk. Baobabview: Interactive construction and analysis of decision trees. 2011 IEEE conference on visual analytics science and technology (VAST). IEEE, 2011.

¹¹ Mitchell, Michael N. Interpreting and visualizing regression models using Stata. Vol. 558. College Station, TX: Stata Press, 2012.

않는다. 이러한 모델의 예측 결과는 학습과 예측이 이루어진 후에 해석하는 것이 유리하다^{1,12}.

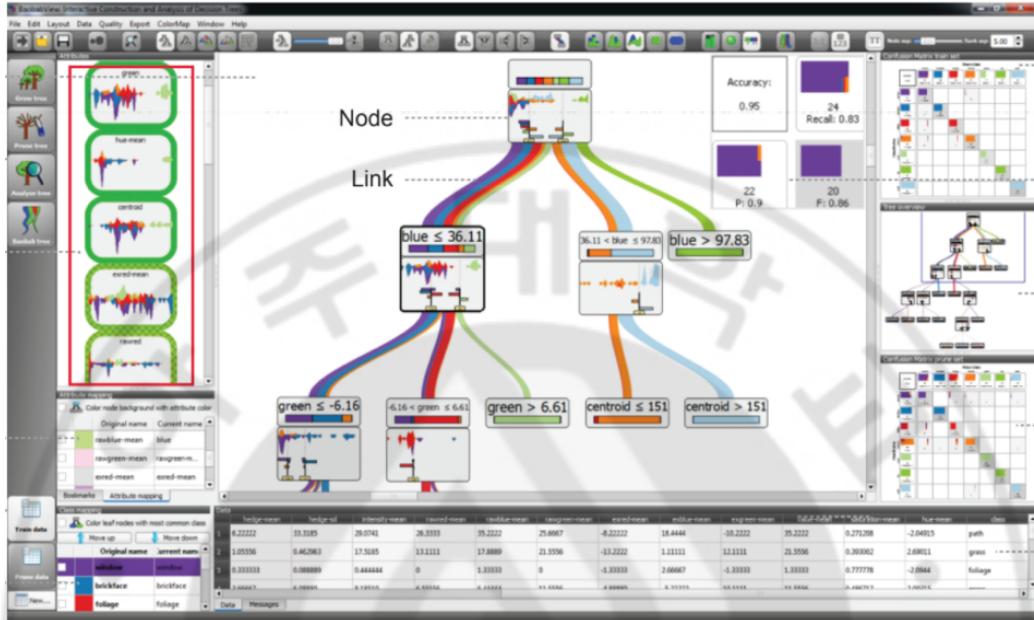


그림 1. 의사결정나무의 예측 과정을 시각화한 BaobabView¹⁰. 의사결정나무는 그것이 가진 계층적 구조 특성 덕분에 시각화 및 해석이 인공 신경망 모델에 비해 수월하다.

¹² Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). International conference on machine learning. PMLR, 2018.

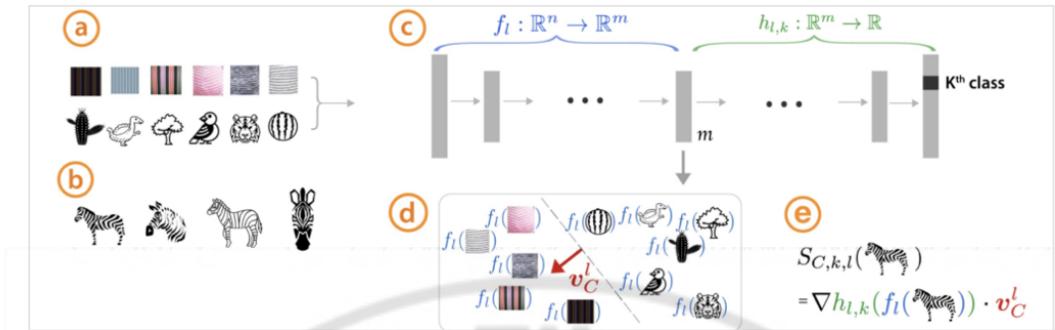


그림 2. 인공 신경망 모델의 의미론적 해석을 지원하는 Testing with Concept Activation Vectors (TCAV) 방법론¹². 모델 구조의 다양성 및 복잡성 때문에 사후 해석이 필요하다.

머신러닝 모델 해석은 기법이 특정한 구조의 모델에만 적용될 수 있는지 여부에 따라 특화된 방식(model specific)과 범용적 방식(model agnostic)으로 분류된다⁹. 모델 특화 방식은 해석 방식이 다양한 모델에 적용 불가능하다는 단점이 있다. 또한 이 방식은 모델 구조에 대한 사전 지식을 필요로 하기 때문에 비 전문가가 사용하기에는 부적합하다¹³. 모델 해석 결과를 모델 연구자 뿐만 아니라 고객이나 협업자와 같은 머신러닝 비 전문가와 함께 공유해야 하는 상황에서 이러한 단점이 부각된다¹. 이러한 경우에 모델 내부에 접근하지 않고 모델을 해석하는 범용적 방식을 제공하는 것이 대안이 된다¹⁴.

¹³ Wang, Z. J., Turko, R., Shaikh, O., Park, H., Das, N., Hohman, F., ... & Chau, D. H. P. CNN explainer: Learning convolutional neural networks with interactive visualization. IEEE Transactions on Visualization and Computer Graphics 27.2, 2020.

¹⁴ Ren, D., Amershi, S., Lee, B., Suh, J., & Williams, J. D.. Squares: Supporting interactive performance analysis for multiclass classifiers. IEEE transactions on visualization and computer graphics 23.1, 2016.

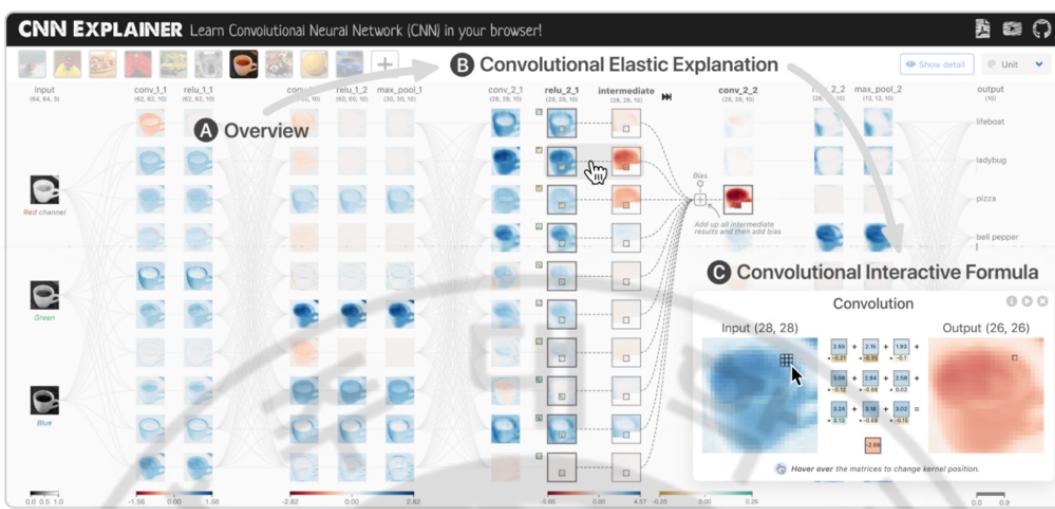


그림 3. CNN 모델의 구조적 특성을 시각화하여 모델의 작동 원리 및 예측 결과를 해석한 CNN explainer¹³. 이 시스템은 시각 분석 요소에 CNN 모델의 구조를 사용했기 때문에 이에 대한 이해가 부족하다면 사용이 어렵다.

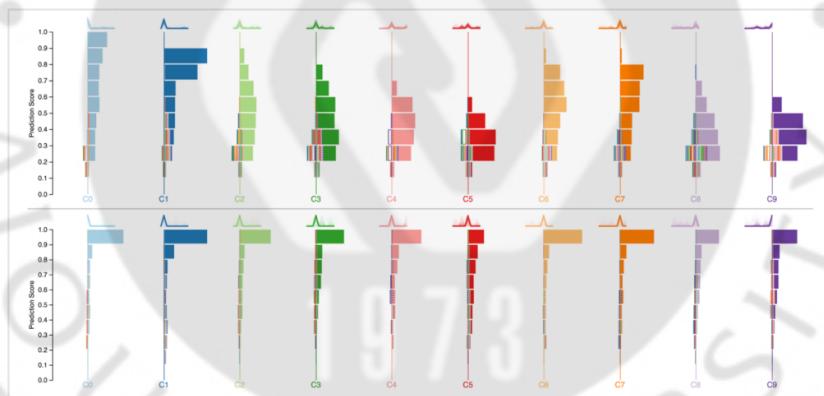


그림 4. 모델 예측의 정오를 클래스별로 시각화하여 모델이 혼동하는 클래스를 탐색하는 시각화 Squares¹⁴. 이 시각화는 모델 내부 데이터에 접근하는 대신 모델 예측 결과만을 활용하기 때문에 사용자는 모델 구조에 대한 이해가 필요하지 않다.

머신러닝 모델 해석 기법은 설명 범위에 따라 분류되기도 한다¹⁵. 해석 기법이 모델 전체 예측에 적용 가능할 경우 그 기법을 전역적(global)이라고 말하고, 특정 상황에만 적용 가능할 경우 국소적(local)이라고 말한다. 전역적 기법은 모델의 전반적인 개요를 파악하는 데 적절한 대신 특정 상황에 대한 구체적인 분석이 어렵다는 단점을 지니며, 국소적 기법은 그 반대의 특징을 가진다. 최근 등장하는 인터랙티브 분석 시스템은 모델 해석을 위해 단일한 기법을 제공하는 대신에 다양한 기법을 제공하여 사용자의 분석 자유도를 높인다^{16,17,18}.

¹⁵ Molnar, Christoph, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges. Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2020.

¹⁶ Park, H., Nam, Y., Kim, J. H., & Choo, J., HyperTendril: Visual Analytics for User-Driven Hyperparameter Optimization of Deep Neural Networks, IEEE Transactions on Visualization and Computer Graphics, 27(2), 2020.

¹⁷ Chanhee Park, Hyojin Kim, and Kyungwon Lee. A Visualization System for Performance Analysis of Image Classification Models. Electronic Imaging 2020.1, 2020.

¹⁸ Zhang, J., Wang, Y., Molino, P., Li, L., & Ebert, D. S. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. IEEE transactions on visualization and computer graphics 25.1, 2018.

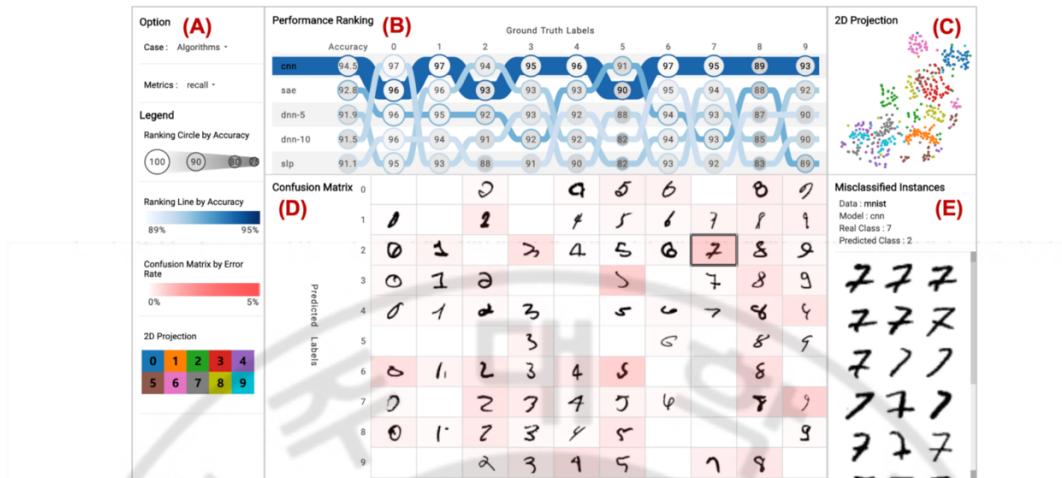


그림 5. 구역 구분을 통해 모델(B, C), 클래스(D), 인스턴스(E) 단위의 해석을 각각 제공하여 전역적 분석과 국소적 분석을 모두 가능하도록 지원하는 인터랙티브 시각 분석 시스템 사례¹⁷.

머신러닝 모델 해석 기법은 제공받는 사람에 따라 다르게 설계되어야 한다¹⁹. 해석 기법의 사용자는 머신러닝 실무자, 데이터 분석가, 비 전문가로 분류될 수 있다. 머신러닝 실무자는 모델을 만들고 수정하는 사람, 즉 모델 개발자이다. 이들은 모델 작동을 이해하고, 오분류 원인을 파악하고자 하며, 모델에 대한 디버깅 기능을 필요로 한다. 이들을 위한 시스템은 데이터, 모델 구조, 변수의 변화가 성능에 미치는 영향을 보는 데 초점을 맞춘다^{13,20,21}. 데이터 과학자 및 분석가는 모델이 내놓은 결과 사용하는 사람, 즉 모델 사용자이다. 이들은 결과를 실무에 적용하기 전에 해당 결과가 추론된

¹⁹ Park, Y., and J. Y. Yun. A Design Case Study of Artificial Intelligence Pipeline Visualization. Archives of Design, 2021.

²⁰ Ming, Yao, et al. Understanding Hidden Memories of Recurrent Neural Networks, 2017 IEEE Conference on Visual Analytics Science and Technology (VAST). IEEE, 2017.

²¹ Kahng, M., Andrews, P. Y., Kalro, A., & Chau, D. H., ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models, EEE transactions on visualization and computer graphics 24.1, 2017.

근거를 해석할 필요가 있다. 이들을 위한 시스템은 모델 예측 결과를 해석하는 데 도움을 주는데 목적을 맞춘다. 이 분류의 연구는 데이터와 예측 결과 사이의 관계를 보여주는 게 중요하다^{12,22}. 비 전문가는 개별 머신러닝 모델보다는 머신러닝 기술 자체를 이해할 필요가 있는 학습자다. 이를 돋기 위한 연구는 모델 내부 구조 및 가중치를 시각화하여 모델 작동 원리를 설명한다^{23,24}.



그림 6. 하이퍼파라미터 변화가 모델 성능에 주는 영향을 중심으로 모델을 시각화한 HyperTendril¹⁶. 모델 연구자 및 개발자들은 모델 내부 정보인 하이퍼파라미터를 중심으로 모델을 이해하고 해석하는 기법을 필요로 한다.

²² Ming, Yao, Huamin Qu, and Enrico Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics* 25(1), 2018.

²³ Smilkov, D., Carter, S., Sculley, D., Viégas, F. B., & Wattenberg, M., Direct–Manipulation Visualization of Deep Networks. *Direct-manipulation visualization of deep networks*. arXiv preprint arXiv:1708.03788, 2017.

²⁴ Minsuk Kahng, Nikhil Thorat, Polo Chau, Fernanda Viégas, and Martin Wattenberg. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. *IEEE Transactions on Visualization and Computer Graphics*, 25(1) (VAST 2018), Jan. 2019.

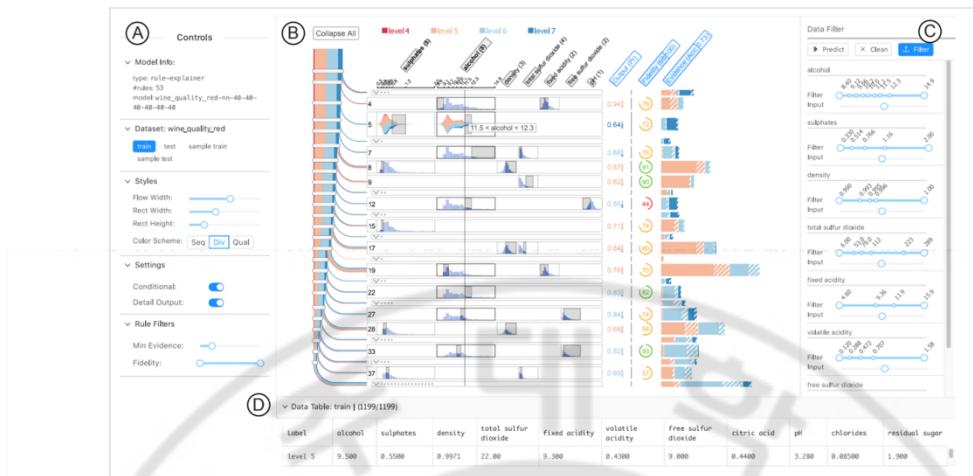


그림 7. 인공 신경망 모델을 근사하는 단순한 구조의 규칙 기반 모델을 생성하여 이를 시각화하는 RuleMatrix²². 데이터 분석가 및 모델 사용자에게는 데이터의 특성과 모델의 예측 결과 사이의 특성을 연결지어주는 기능이 필요하다.

최근 들어 머신러닝 분야에는 복잡한 문제를 해결하기 위해 다양한 모델이 도입되고 있다. 또한, 머신러닝 모델 사용자는 머신러닝 분야에 대한 사전 지식이 적을 수 있다. 이러한 경우에는 단순한 모델을 이용하거나, 모델에 특화된 방식을 사용하여 모델을 해석하는 것이 부적합하다. 본 연구는 머신러닝 모델을 쉽게 해석할 수 있도록 지원하는 시각 분석 시스템을 설계하고 개발하는 것을 목표로 한다. 이에 따라, 본 연구는 모델이 내놓은 결과를 활용하는 이들에게 적합한 방식인 모델 구조에 관계없이 적용 가능하고, 예측 결과를 바탕으로 한 해석을 제공하는 방식에 초점을 맞춘다. 또한, 본 연구는 시각 분석 시스템 설계에 필터링 및 자세히 보기 등의 인터랙션을 도입하여 모델 예측 결과를 전역적 차원과 국소적 차원 모두에서 분석할 수 있도록 한다.

제 2 절 입력 변수 값 변형을 활용한 모델 해석

본 연구에서는 머신러닝 모델 해석을 위한 시스템 설계에 입력 변수 값을 변형시키고 변형 전후의 모델 출력 값을 비교하는 기법을 도입하였다. 분류와 예측 문제에서 머신러닝 모델은 입력 변수를 바탕으로 타깃 변수의 값을 예측한다. 직관적 모델 해석 방법 중 하나는 입력 변수 값 변화에 따른 예측 값의 차이를 살펴보는 것이다. 최근에는 입력 변수 값 변화 전후로 모델 예측 값이 어떻게 달라지는지 분석해 모델을 해석하는 연구가 진행되어왔다.

Partial Dependence Plot(부분 의존도 그래프, 이하 PD 플롯으로 표기)은 개별 입력 값이 모델의 예측에 미치는 한계 효과(marginal effect)를 보여준다²⁵. 이 방식은 검사를 원하는 입력 변수의 값을 변경한 뒤에 타깃 변수의 값이 어떻게 변화했는지 평균을 내어 그래프를 그린다. 그 그래프에는 타깃 변수와 입력 값 간의 관계가 선형인지, 단조로운지 또는 더 복잡한지 여부가 나타난다. 입력 변수 값 변경이 타깃 변수 예측 값 변화에 어떤 영향을 주는지 보여주는 PD 플롯의 전략은 직관적이다. 다만, PD 플롯은 최대 두 개의 입력 변수에 대해서만 모델의 효과를 확인할 수 있다는 단점이 있다. 이러한 한계를 지니는 이유는 PD 플롯이 변경된 입력 변수와 타깃 변수를 각각 축으로 가져야 하는 데 3 차원 이상의 공간을 시각적으로 표현할 수 없기 때문이다. 또한, PD 플롯은 모든 변화된 인스턴스의 평균값을 바탕으로 모델을 해석하는 점에서 한계를 지닌다. PD 플롯의 주된 시나리오는 선택된 입력 변수 X 가 변화할 때 타깃 변수 Y 예측 값이 어떻게 바뀌는지 보는 것이다. 이때, 인스턴스는 두 그룹(예를 들어, 남성과 여성)으로 나뉠 수 있다. 한 그룹에서는 X 가 Y 에 긍정적 영향을 주고, 다른 그룹에서는 X 가 Y 에 부정적

²⁵ Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001.

영향을 준다면, 전체 인스턴스에서는 X 변화는 Y 변화 평균을 변화시키지 못한다. 이런 경우에 PD 플롯은 X의 영향을 감지할 수 없다.

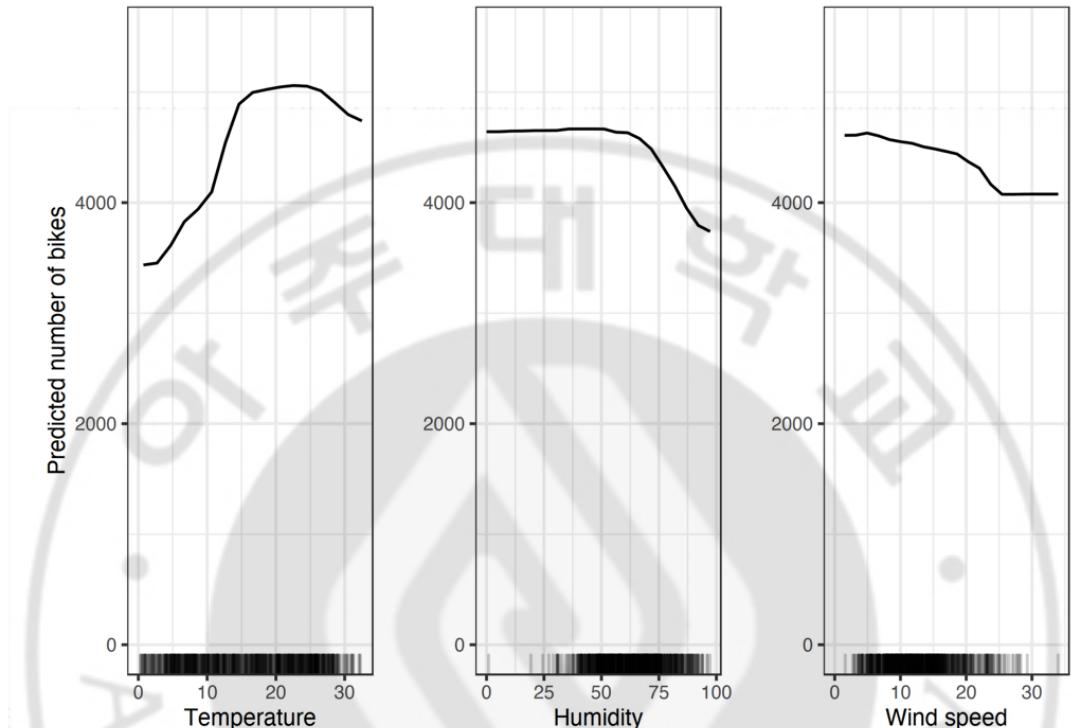


그림 8. 입력 값의 변화에 따른 예측 값 변화 추세를 보여주는 PD Plots. 각 PD Plot 은 개별 입력 변수 값 변화에 따른 예측 값 변화만 보여줄 뿐 동시에 여러 입력 변수 값이 변한 상황을 보여주지 못한다.

PD 플롯이 평균값을 사용하기 때문에 생기는 문제를 방지하는 방법은 개별 인스턴스의 변화를 살펴보는 것이다. Individual Conditional Expectation Plot(개별 조건부 기대치 플롯, 이하 ICE 플롯으로 표기)은 입력 변수가 변경될 때 예측이 어떻게

변하는지 보여주는 선을 인스턴스 당 하나씩 표시한다²⁶. 선의 값은 선택된 입력 변수 하나를 새로운 값으로 변경하여 변형된 인스턴스에 대한 타깃변수 예측 값을 표시한다. ICE 플롯은 PD 플롯보다 더 직관적으로 이해할 수 있다. 또한 PD 플롯과 달리 ICE 플롯의 곡선은 평균 대신 각각의 인스턴스 변화 값을 표현하기 때문에 이기종 관계(heterogeneous relationships)를 발견할 수 있다. 인스턴스를 선택하여 해당 인스턴스의 다른 입력 변수가 변화되었을 때의 효과도 파악할 수 있다는 점에서 인터랙션의 강점도 가진다. PD 플롯과 ICE 플롯을 각기 다른 색상으로 중첩하여 표현한다면 양쪽의 장점을 모두 누릴 수 있다. 하지만 ICE 플롯은 최대 한 개의 입력 변수 변화에 대해서만 시각화할 수 있고, 인스턴스 수가 많아지면 너무 많은 선이 과밀화되어 알아볼 수 없으며, 인스턴스 별 변화 양상이 다양해진다면 그 그룹별 차이를 보는 것이 어려워지는 단점을 지닌다.

²⁶ Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E., Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24.1, 2015.

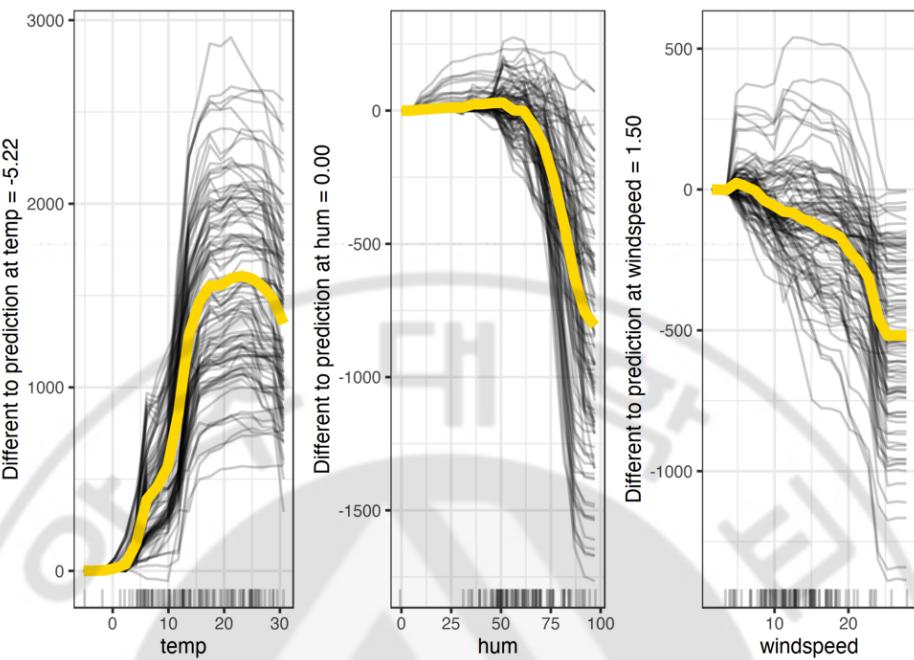


그림 9. 입력 값의 변화에 따른 개별 인스턴스의 예측 변화를 보여주는 ICE Plots. 노란색 선은 모든 인스턴스의 평균 값을 표시한 것이다. 인스턴스가 많아질 경우 시각화 선이 얹혀 추세를 파악하기 어려워진다.

DECE는 이진 분류 모델의 결정을 해석하기 위한 시각 분석 시스템이다²⁷. 이 시스템은 사용자가 다양한 가설을 실험할 수 있도록 유연하게 인스턴스 그룹을 만들어 해석 결과를 비교할 수 있도록 돋는다. 또한 이것은 선택된 인스턴스의 특정 입력 변수가 어떤 값으로 바뀌어야지 예측 값이 변화하는지 보여준다. 이러한 과정은 사용자가 머신러닝 모델의 해석을 통해 입력 변수와 목적 변수 사이의 관계에 대한 추론을 유도한다. 또한 이것은 히스토그램을 활용해 쉽게 인스턴스 그룹을 만들 수 있도록 유도한다는 측면에서 가치를 지닌다. 하지만 이 히스토그램은 데이터의 분포만을 보여줄

²⁷ Cheng, Furui, Yao Ming, and Huamin Qu. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. IEEE Transactions on Visualization and Computer Graphics 27.2, 2020.

뿐, 해당 분포가 예측하려는 클래스와 어떤 관련이 있는지 보여주지 않기 때문에 사용자는 어떤 입력 변수를 위주로 탐색해야 할지에 대한 정보가 부족한 상태에서 그루핑과 필터링을 진행해야 한다. 또한, 이 DECE 는 모델 예측의 정오를 테이블 시각화의 첫 번째 열에서만 보여줄 뿐이다. 따라서, 이 시스템은 모델 예측 정오에 대한 인사이트를 적게 반영하고, 그저 입력 변수와 타깃 클래스 사이의 관계만을 보여준다는 지적을 피하기 어렵다. 결정적으로, 이 시스템은 다중 분류나 회귀 문제에 적용되지 않는다는 단점이 있다. 다양한 종류의 모델을 해석하기 위한 테이블 시각화 디자인은 여전히 도전과제로 남아있다.

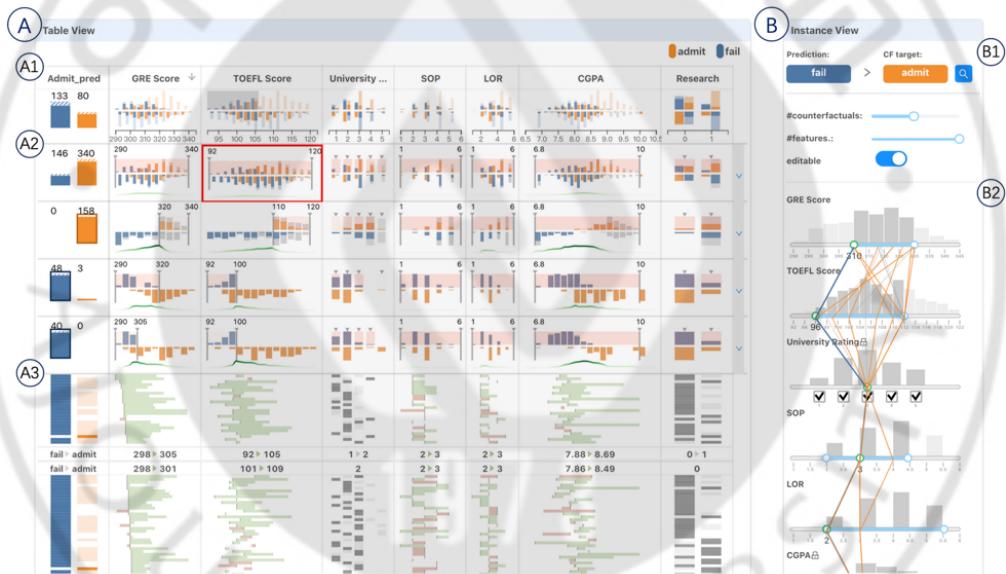


그림 10. 인스턴스를 그룹화하여 입력 변수와 모델 예측 사이에 상관관계가 존재하는지 파악하도록 돋는 DECE. 이진분류 모델에만 작동가능하며 입력 변수들 사이 상관관계를 파악하기 어렵다는 한계가 존재한다.

입력 데이터 변형과 이에 따른 모델 출력 값 변화 추적 연구들은 모델이 입력 변수와 예측 변수 사이의 관계를 어떻게 연결 짓는지를 보여주어 모델의 작동 패턴을 해석해왔다. 한편 이들 연구는 데이터가 분포한 전반적인 패턴을 파악하기 어렵고,

인스턴스 전체와 개별 인스턴스를 자유롭게 오가며 분석을 수행할 수 없다는 한계를 가졌다. 이를 통해, 본 연구는 모델 해석을 위해 입력 데이터 값이 변형된 데이터를 활용하는 방안의 가치와 도전 과제를 발견했다.

제 3 장 연구의 목적과 방법

본 연구의 목적은 머신러닝에 전문 지식이 부족한 실무자가 머신러닝 모델을 사용함에 있어서 사용자가 모델을 해석하고 이해할 수 있도록 돋는 시각 분석 시스템을 만드는 것이다. 본 연구는 이와 같은 목적을 달성하기 위하여 모델 평가에 사용된 데이터 셋을 다양한 형태로 변형시켜 모델의 결정을 다각도로 분석하는 시스템을 제안한다. 특히, 입력 변수 변화가 모델 예측에 어떠한 변화를 만들어 내는지를 보여주기 위한 시각화와 인터랙션 방법에 연구 초점을 맞춘다.

변형된 데이터를 바탕으로 모델을 해석하는 연구들은 입력 변수 값 변화 전후로 모델 예측 값이 어떻게 달라지는지 분석한다. 이 덕분에 사용자는 개별 입력 변수가 모델 예측에 미치는 한계 효과를 직관적으로 분석할 수 있다는 장점이 있다. 하지만, 이러한 기여에도 불구하고 기존 연구들은 다음과 같은 한계점을 보였다. 첫째, 데이터 변형에 조건이 걸려있지 않기 때문에 현실성이 떨어지는 데이터로 변형될 수 있다. 기존의 연구들은 모든 입력 변수에 대한 모든 가능한 값을 도입하거나, 임의 입력 변수에 대해 임의 값을 도입하여 데이터 변형을 시도해왔다. 이러한 방식은 데이터 사이의 상관관계를 무시하여 현실성이 떨어지는 데이터(예를 들어, 키가 200cm이고, 몸무게가 30kg인 사람)를 많이 만들어낸다. 현실에 존재하지 않을 법한 데이터를 바탕으로 모델을 해석하는 것은 모델의 약점을 찾는 데에는 도움이 될 수 있지만, 일반적인 상황에서의 모델의 행동양식을 살펴보는 데에는 부적합하다.

SMOTE 알고리즘은 데이터 오버샘플링을 위한 기법 중 하나이다²⁸. 이 기법은 KNN (k-최근접 이웃 알고리즘)을 바탕으로 개별 인스턴스와 가장 유사한 인스턴스를 찾아 그 두 인스턴스 사이 값에 해당하는 값으로 새로운 인스턴스를 생성한다. 이 기법을 활용하면 현실 세계에서 존재할 법한 인스턴스를 만들 수 있다. 본 연구는 SMOTE 알고리즘을 모델 해석을 위한 데이터셋 변형 시점에 적용시켜 분석 시스템이 보다 현실에 가까운 분석 결과를 도출하도록 만들었다.

SMOTE 알고리즘을 활용하여 인스턴스의 입력 변수 값을 변화한 방법은 다음과 같다. 첫째, 하나의 인스턴스(원본 인스턴스)를 골라 해당 인스턴스와 입력 변수 값의 분포가 유사한 인스턴스 10 개(변형 조건 인스턴스)를 선택한다. 이때, 유사성 판단에는 맨해튼 거리 알고리즘을 사용한다. 둘째, 원본 인스턴스에서 변형할 변수와 참조할 변형 조건 인스턴스 하나를 선택하여 변수 값 변형을 진행한다. 만약 원본 인스턴스가 29 살 남성이고 변형 조건 인스턴스가 30 살 여성인데 변형할 변수가 성별이라면, 변형된 인스턴스는 29 살(원본 인스턴스 값 참조) 여성(변형 조건 인스턴스 값 참조)이 된다. 이 과정을 모든 가능한 두 개 이하의 변수 선택 조합에 대하여 반복한다. 이 과정을 변형 조건 인스턴스 10 개에서 반복한다. 마지막으로 원본 인스턴스로 모든 인스턴스가 활용될 때까지 위 작업을 반복한다. 이 과정을 통해 모든 가능한 두 개 이하의 변수 조합에 대하여 합리적인 값 범위 내에서의 변수 값 변형이 이뤄지게 된다. 인스턴스의 수가 N 개 변수의 수가 M 개라고 할 때, 총 $N * 10 * (({}_M C_2) + ({}_M C_1))$ 개의 변형 인스턴스가 생성된다.

²⁸ Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., SMOTE: synthetic minority over-sampling technique. Journal of artificial intelligence research 16, 2002.

데이터 변형에 조건이 걸려있지 않다는 것 외에 기존 연구들의 또 다른 한계점은 사용자에게 해석의 단서를 충분히 제공하는 대신 단순히 모든 가능한 결과를 나열한다는 점이다. 머신러닝 모델이 해결하고자 하는 문제는 점점 복잡해지고 있다. 기존의 분석 툴이 제공하는 해석 방식은 입력 변수와 인스턴스 수가 많아질 때 사용이 어렵다. 또한, 머신러닝 전문 지식이 부족한 일반 실무자는 실제 현상과 모델 예측 결과, 또는 실제 데이터와 생성된 데이터를 구분 지어 해석하기 어려워할 수 있다²⁹. 이에 따라, 본 연구는 머신러닝 전문 지식이 부족한 일반 실무자가 머신러닝 모델을 이해하고 해석하는 데 드는 어려움을 덜기 위하여 시각화 요소와 인터랙션 기법을 활용하여 유저 인터페이스 기반 분석 시스템을 설계하고자 한다.

Schneiderman 에 의해 창안된 시각화 만트라는 데이터 분석 시스템 설계 방침 중 하나로, 방대한 데이터로부터 주요 발견점을 찾을 수 있도록 돋는다³⁰. 시각화 만트라는 사용자가 (1) Overview, (2) Filter & Sorting, (3) Details 를 순서로 데이터를 분석하도록 유도한다. 본 연구는 사용자에게 해석의 단서를 충분히 제공하기 위하여 이 3 단계의 분석 순서를 유도한다. 다만 이 만트라는 관심 범위를 좁히는 단계에서 필터 기능만 사용한다는 한계가 있다. 최근 등장하는 인터랙티브 분석 시스템은 필터뿐만 아니라 그룹, 정렬, 특정 값 찾기, 입력 변수 간의 상관관계 제공하기 등 다양한 분석 기법을 제공하여 사용자의 분석 자유도를 높였다^{16,17,18,31}.

²⁹ Kliegr, Tomáš, Štěpán Bahník, and Johannes Fürnkranz. A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artificial Intelligence*, 2021.

³⁰ Schneiderman, Ben. *The eyes have it: A task by data type taxonomy for information visualizations. The craft of information visualization*. Morgan Kaufmann, 2003.

³¹ Amar, Robert, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. *IEEE Symposium on Information Visualization*, 2005.

특히, 머신러닝 모델 해석이 어려워지는 이유는 인스턴스 수와 입력 변수의 수가 늘어가고 해결하고자 하는 문제가 복잡해짐에 따라 개별 인스턴스의 특징이 전체 데이터 셋을 대표하기 어려워진 데에 있다. 이에 따라 여러 인스턴스를 한 번에 살펴볼 수 있는 기능의 필요성이 대두되고 있다. 본 연구는 필터에, 그룹, 정렬 기능을 중심으로 통해 사용자가 보다 쉽고 빠르게 모델 해석에 대한 발견점을 얻을 수 있도록 도왔다.

본 연구는 입력 데이터 변화가 모델 결정에 미치는 영향을 추적하여 사용자가 모델을 해석하는 것을 돋는 것에 목표를 두고 시각 분석 시스템을 디자인하고 평가한다. 과학적 이해, 안정성 파악, 신뢰 확보와 같은 해석 가능한 머신러닝의 장점을 획득할 수 있는지가 시스템 디자인의 주요 고려사항이다⁹. 구체적으로 본 연구가 제안하는 시각 분석 시스템은 다음과 같은 4 가지 디자인 목표를 달성하고자 한다 (G1 ~ G4).

- G1: 입력 변수와 모델 예측 값 사이의 관계 파악. 각 입력 변수 별로 머신러닝 모델의 예측에 영향을 주는 상대적인 정도를 파악할 수 있다. 이 목표를 통해 사용자는 특정 머신러닝 모델이 어떤 입력 변수에 민감하거나 민감하지 않은지 파악할 수 있다. 모델의 예측 정확도가 높은 경우에는 모델이 학습한 바를 해석하여 입력 변수와 타깃 변수 사이의 관계에 대한 지식을 얻을 수 있다. 또한 인종과 성별 등 적용되지 않아야 할 편견적 요소가 모델에 학습되어 있는지 파악할 수 있다.
- G2: 입력 변수와 모델 예측 오류 사이의 관계 파악. 머신러닝 모델의 예측 정확도를 높이는 입력 변수 목록과 오류를 크게 만드는 입력 변수 목록을 구할 수 있다. 이 목표가 달성될 경우 사용자는 머신러닝 모델의 행동과 예측 패턴을 이해하고 결정을 신뢰하거나, 오류를 파악하고 대응할 수 있게 된다.
- G3: 입력 변수 변화에 따른 모델 예측 변화 추이 파악. 특정 입력 변수 및 변수의 조합의 값 변화가 모델 예측의 변화를 크게 만드는지 파악한다. 이 목표의 달성을 통해,

사용자가 모델이 작동하는 방식에 대해 이해하도록 도울 수 있다. 또한 모델 예측 정확도가 낮은 경우에는 작은 입력 값 변화가 큰 예측 오류로 이어지는 상황을 파악하여 모델의 안정성을 진단할 수 있다.

- G4: 개별 인스턴스 확인. 유저가 분석하는 대상의 개별 인스턴스 목록을 파악할 수 있다. 개별 인스턴스 목록을 아는 것은 그렇지 않을 때 보다 분석 결과를 스토리로 만들고 타인과 공유하는 데 용이하게 만든다³².

제 4 장 시각 분석 시스템 설계 및 개발



그림 11. 연구 결과인 시각 분석 시스템을 통해 머신러닝 모델을 분석하는 과정

본 연구가 제안하는 시스템이 작동하는 개략적인 과정은 다음과 같다. 우선, 유저는 데이터 및 모델을 시스템에 입력한다. 이 시스템이 다루는 데이터의 유형은 정형 다차원 데이터이며, 모델은 분류 또는 수치 예측 문제를 해결한다. 데이터는 CSV 파일로 입력하며 예측하고자 하는 타깃 변수의 이름은 ‘actual value’로 지정해야 한다. 데이터와 모델이 입력되면 시스템은 데이터 전처리를 수행한다. 이 과정에서 모델은 결측치를 제거하고, 기존 데이터를 바탕으로 입력 데이터를 변형한 신규 데이터를 생성한다. 다음으로 시스템은 처리된 기존 데이터와 생성된 신규 데이터를 시각화한다. 기존 데이터는 시스템 상단 평행 좌표 그래프와 산점도에 표시되며, 생성된 데이터는

³² Aamodt, Agnar, and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI communications 7.1 1994.

시스템 하단 표에 나타난다. 유저는 시각화와 인터랙션을 활용하여 모델 분석을 수행한다.



그림 12. 본 연구가 제안하는 시각 분석 시스템의 인터페이스. 이 사례는 인적 정보를 바탕으로 의료 비용 지출액을 예상하는 모델을 해석한다. 해석 결과는 5 장 유스 케이스 분석에 작성되어 있다. (A) 영역은 데이터, 모델 정보, 시각화 매핑 기법을 소개한다. (B) 영역은 평행 좌표 그래프로 모델 해석을 위한 실제 인스턴스의 분포를 보여준다. 이곳에서 유저는 인스턴스를 필터링할 수 있다. (C) 영역은 인스턴스의 입력 변수 별 분포를 보여주는 산점도 목록이다. 각 산점도는 입력 변수를 세로축, 모델 예측 값을 가로축으로 표시하여 입력 변수의 값과 모델 예측 값 사이의 관계를 표시한다. (D)는 입력 변수가 변형된 인스턴스를 변형된 입력 변수의 종류 별로 그룹지어 히스토그램으로 시각화한 테이블이다.

제 1 절 평행 좌표 그래프 및 산점도

이 영역에서는 오버뷰 시각화를 통해 입력 변수와 모델 예측 사이의 개요를 보여준다. 인스턴스 및 입력 변수의 수가 많아지는 최근의 머신러닝 실무 및 연구 환경에서 모델 해석을 위해서는 사용자에게 해석의 단서를 충분히 제공해야 한다. 또한 입력 변수 값의 변형을 통한 모델 해석을 달성하기 위해서는 입력 변수의 특성을 이해할 수 있도록

지원해야 한다. 사용자가 학습된 모델과 검증에 사용할 데이터셋을 시스템에 입력하면, 시스템은 입력받은 인스턴스의 모델 예측 값을 계산한다. 예측 결과는 평행 좌표 그래프(Parallel Coordinates Plot) 및 산점도(Scatter Plots)를 통해 시각화된다. 이 기법을 활용하면 입력 변수와 타깃 변수, 예측 결과 사이의 상관관계를 개략적으로 살펴볼 수 있다(G1).

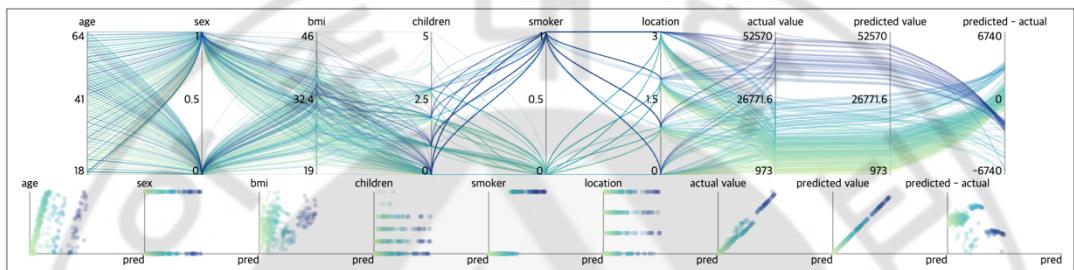


그림 13. 제안된 시스템의 평행 좌표 그래프와 산점도.

평행 좌표 그래프는 입력 변수, 타깃 변수의 실제 값(actual value), 예측 값(predicted value), 타깃 변수의 예측 값과 실제 값 사이의 차이(predicted – actual)를 축으로 가진다. 입력 변수와 관련된 축을 좌측, 타깃 변수와 관련된 축을 우측에 배치하였다. 입력 변수 축의 순서는 사용자가 입력한 CSV 파일의 순서를 따른다. 하나의 선은 하나의 인스턴스를 의미하며, 각 축에서의 선의 높이가 해당 인스턴스가 가지는 값을 표시한다. 평행 좌표 그래프 바로 아래에 위치한 각 산점도는 세로축을 입력 변수, 타깃 변수의 실제 값, 예측 값, 타깃 변수의 실제 값과 예측 값 사이의 차이 값을 가지고, 가로축으로 타깃 변수의 모델 예측 값을 가진다. 가로축은 사용자의 설정에 따라 타깃 변수의 실제 값으로 변경할 수 있다. 산점도에서 하나의 점은 하나의 인스턴스를 의미한다. 평행 좌표 그래프와 산점도 축의 순서와 가로 위치를 같게 하여 사용자가 관심 있는 축을 빠르게 찾을 수 있도록 지원하였다. 또한 평행 좌표 그래프와 산점도 모두에서 인스턴스의 색을

타깃 변수의 값에 매핑시켰다. 사용자는 인스턴스의 색을 통해 타깃 변수의 예측 값을 높게 만드는 인스턴스가 다른 입력 변수에서 가지는 값의 분포를 파악할 수 있다. 수치형 예측의 경우 타깃 변수의 예측 값이 클록 진한 푸른색을 가지고, 낮을수록 연한 녹색으로 표시된다. 분류형 예측의 경우 예측된 값이 참일 경우 푸른색, 거짓일 경우 붉은색을 가지게 하였다. 컬러 스케일 및 색상을 매핑할 변수는 사용자의 설정에 의해 타깃 변수의 실제 값이나 기타 변수 등으로 변경될 수 있다. 이 두 시각화를 통해 사용자는 상관관계가 깊은 변수 목록을 찾는데 도움을 받을 수 있다(G1, G2).

전역적 기법은 모델의 전반적인 개요를 파악하는 데 적절한 대신 특정 상황에 대한 구체적인 분석이 어렵다는 단점을 지닌다. 이러한 단점을 극복하기 위해 시각 분석 시스템에는 전역 환경에서의 개요를 제공하는 한편, 인터랙션을 통해 국소적인 상황으로 심층 분석할 수 있는 기능이 요구된다. 또한, 머신러닝 모델이 입력 변수의 값으로부터 받는 영향을 보기 위해서는 입력 변수와 예측 값을 보는 것에 더해서 입력 값의 변화가 예측 값에 어떤 변화를 주는지 살펴보는 것이 도움이 된다^{25,26}. 전체 인스턴스 입력 값의 변화가 모델 전반의 예측에 주는 영향을 분석하는 방법은 대다수의 평균적 상황에 특정한 상황이 가려질 수 있다는 한계가 발견되었다. 반면, 개별 인스턴스의 입력 값 변화에 대해서 모델 예측의 변화를 살펴보는 연구는 해당 케이스가 얼마나 일반적으로 드러나는지 확인하기 어렵다는 단점을 가졌다³³. 이에 따라, 평행 좌표 그래프와 산점도 영역은 필터, 생성, 그룹, 정렬 인터랙션을 통해 모델 해석을 지원한다.

³³ Shi, Sheng, Xinfeng Zhang, and Wei Fan. A modified perturbed sampling method for local interpretable model-agnostic explanation. arXiv:2002.07434, 2020.

제 2 절 인스턴스 변형과 테이블 시각화

시스템 하단 테이블 시각화는 필터링된 인스턴스를 원본으로 가지고 있는 변형 인스턴스만을 보여준다. 또한 변형된 데이터 셋을 변형 조건에 따라 그룹 지어 다수의 상황에 의해 나머지 특수한 경우가 가려지지 않도록 했다. 뿐만 아니라, 해당 그룹들을 예측 값 변화량이 큰 순서로 정렬시켜 사용자가 먼저 살펴봐야 할 그룹이 무엇인지 안내했다. 평행 좌표 그래프와 산점도가 모델이 여러 입력 변수에 어떻게 반응하는지 전역적인 개요를 보여줬다면, 테이블 시각화는 유저 인터랙션을 통해 특정 상황에 해당하는 데이터를 선택하여 국소적인 해석을 지원한다(G3). 인스턴스 변형은 관심사를 좁히는 필터, 해당 필터에 해당하는 인스턴스를 변형하여 테스트 셋을 만드는 생성, 생성된 인스턴스를 유형에 따라 나누는 그룹, 그룹들을 중요도 순으로 줄짓는 정렬의 순서로 진행된다.

우선, 필터 단계는 평행 좌표 그래프와 산점도를 통해 수행된다. 이 두 시각화는 드래그 인터랙션을 제공하여 인스턴스를 선택할 수 있게 만든다. 사용자는 평행 좌표 그래프와 산점도의 드래그 인터랙션을 바탕으로 관심 범위를 좁힌다. 예를 들어, 예측과 실제값이 크게 다른 케이스, 예측 값이 극단적으로 높거나 낮은 케이스 등을 드래그로 필터링할 수 있다.

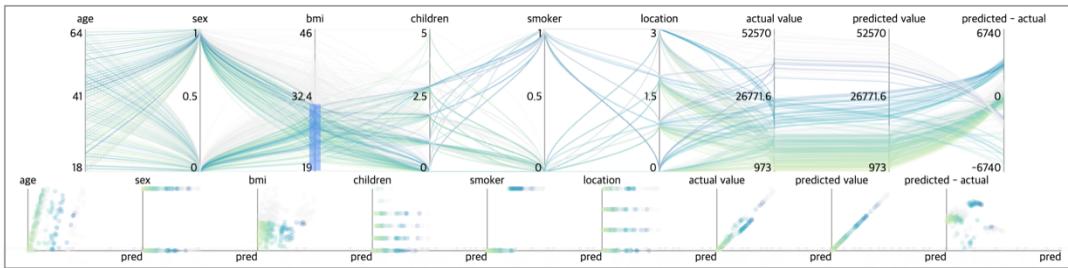


그림 14. 필터가 적용된 평행 좌표 그래프와 산점도. 평행 좌표 그래프의 축에서 마우스 드래그를 통해 인스턴스를 선택할 수 있다. 위 그림에서는 BMI 지수가 19 이상 32 이하인 그룹이 선택되었다. 인스턴스는 유채색, 선택되지 않은 인스턴스는 밝은 무채색으로 표시된다.

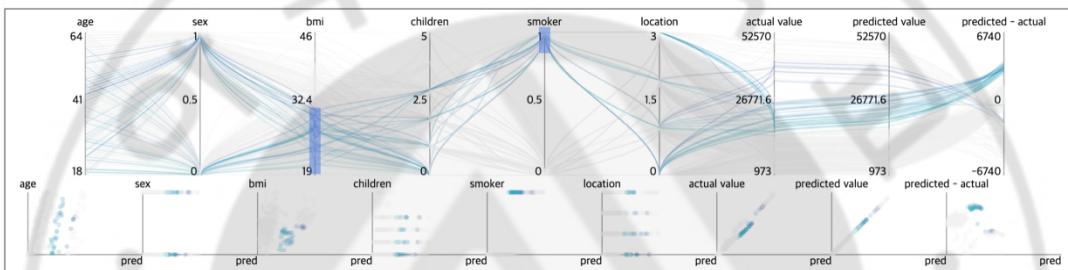


그림 15. 다중 필터가 적용된 평행 좌표 그래프와 산점도. 여러 축에 필터가 지정될 경우, 해당 필터를 모두 만족시키는 인스턴스가 선택된다. 위 그림에서는 BMI 지수가 19 이상 32 이하이면서 smoker(흡연 여부)가 참인 그룹이 선택되었다.

생성 단계는 시스템에 의해 자동으로 수행된다. 시스템은 SMOTE 알고리즘을 사용하여 인스턴스의 입력 변수 값은 변경하여 새로운 인스턴스를 생성한다. 생성된 인스턴스는 어떤 원본 인스턴스로부터 변형되었는지에 대한 정보를 가지고 있다. 이에 따라, 시스템 상단의 평행 좌표 그래프에서 인스턴스 필터링이 이뤄지면 생성된 인스턴스도 함께 필터링된다.

이어서, 생성된 인스턴스를 변경된 입력 변수의 종류 및 변화의 방향을 바탕으로 인스턴스를 그룹 짓는다. 변화의 방향은 세 가지로 나뉘는데, 1) 수치형 변수의 값이 커짐, 2) 수치형 변수의 값이 작아짐, 3) 범주형 변수의 값이 변경됨이 있다. 이 그룹은

테이블 형태로 시각화된다. 테이블 시각화에서 하나의 열(column)은 개별 입력 변수, 타깃 변수의 실제 값, 기존 데이터의 예측 값, 입력 값이 변형된 후의 예측 값을 보여준다. 테이블에서 각 행(row)은 하나의 그룹을 의미한다. 각 그룹은 입력 값이 변형된 인스턴스들로 구성되어 있다.

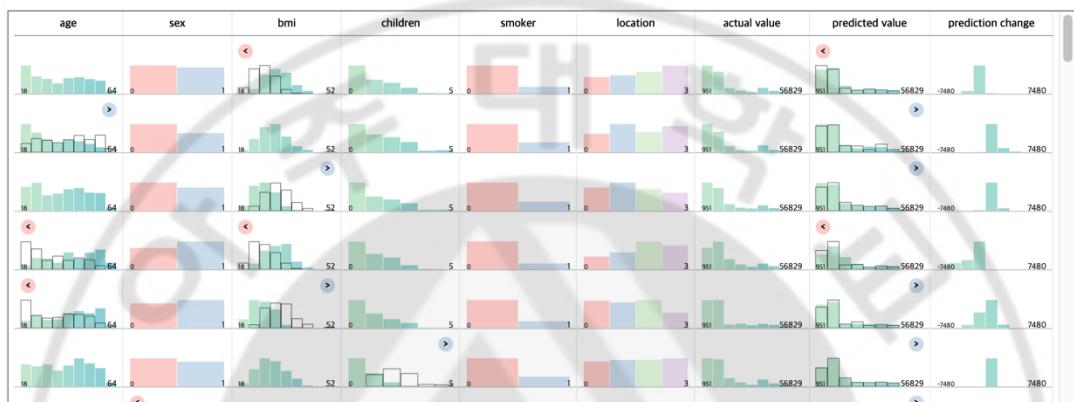


그림 16. 제안된 시스템의 테이블 시각화. 여기에서 각 열은 입력 변수들, 타깃 변수의 실제 값(actual value), 타깃 변수의 모델 예측 값(predicted value), 입력 변수 값 변화 전후 예측된 값의 변화량을 의미한다. 그리고 각 행은 생성된 인스턴스를 변경된 입력 변수의 종류와 변화 방향에 따라 나눈 그룹을 표시한다.

테이블에서 각 셀에는 해당 행의 그룹에 속하는 인스턴스가 해당 열에 해당하는 입력 변수에서 가지는 값의 분포를 표시하는 히스토그램이 시각화된다. 변형이 이뤄지지 않은 입력 변수는 일반 색채 히스토그램만으로 표시된다. 수치형 변수는 값이 클수록 점점 진하게 표시되는 녹색 조 컬러 스케일을 활용했으며, 명목형 변수는 서로 쉽게 구분되는 색상으로 구성된 컬러 팔레트를 사용했다. 사용자는 색상을 이용하여 해당 그룹이 특정 변수에서 어떤 값을 많이 갖는지 빠르게 파악할 수 있다.

변형이 이뤄진 입력 변수는 색채가 있는 히스토그램과 테두리 선으로만 그려진 히스토그램 둘이 합쳐져 표시된다. 이때, 선으로만 표시된 히스토그램은 기존 변수 값의

분포를 표시하며, 색채가 있는 히스토그램이 변형된 값의 분포를 표시한다. 또한 변형이 생긴 입력 변수가 수치형 변수일 때는 셀 상단에는 동그란 칩 디자인 안에 화살표를 표시하여 입력 변수가 전반적으로 상승했는지, 또는 감소했는지 표시한다. 상승한 경우에는 푸른색 동그라미 안에 오른쪽 화살표가 표시되며, 감소한 경우 붉은색 동그라미 안에 왼쪽 화살표가 표시된다.

이 테이블 시각화는 입력 변수 값, 실제 값, 예측 값, 입력 변수의 변화 전후 예측 값의 변화량, 등을 바탕으로 정렬할 수 있다. 정렬된 테이블을 통해 사용자는 입력 변수 값의 변화가 모델 예측 값 변화에 어떤 영향을 주는지 파악할 수 있다 (G3).

age	sex	bmi	children	smoker	location	actual value	predicted value†	prediction change
18 64 0	0 1	18 52 0	52 0 1	0 5 0	1 0	3 991 56829 951 56829	748	Group info: bmi+/smoker+/pred+ -7480

그림 17. BMI 가 높아지고, 흡연 여부가 거짓에서 참으로 바뀐 인스턴스 그룹을 표시한 테이블의 행. 이 그룹은 입력 변수 변경 전보다 모델의 예측 값이 높아졌다. 이러한 변화가 모델의 의료 비용 지출 예측 값을 높게 만드는 경향이 있음이 나타난다. 유저가 테이블의 행 위에 마우스 커서를 올리면 위와 같이 툴 팁 박스가 표시된다. 툴 팁 박스에는 변경된 입력 변수 및 예측 값 정보가 bmi+/smoker+/pred+ (BMI 상승, 흡연 여부 양으로 변경, 예측 값 상승)이 표시되고 있다.

제 3 절 개별 인스턴스 확인

사용자는 “자세히 보기 팝업 창”을 통해 개별 인스턴스 사례를 통해 사례를 탐색할 수 있다. 테이블 시각화는 개별 인스턴스의 값이 아니라 인스턴스 그룹의 통계치를 보여준다. 사용자는 개별 인스턴스 값을 보기 위해 더 자세히 보고 싶은 그룹을 선택할 수 있다. 마우스 클릭을 통해 그룹을 선택하면 팝업 창을 통해 개별 인스턴스 데이터를 보여준다. 이곳에서 사용자는 개별 인스턴스의 입력 변수 값과 예측 값의 변화를 확인할

수 있다. 사용자는 개별 인스턴스 목록을 활용하여 분석 결과를 타인에게 공유하는 데 사용할 수 있다(G4).



그림 18. 테이블 시각화에서 행을 클릭하면 표시되는 해당 그룹에 속한 인스턴스 목록을 보여주는 팝업 창.

제 5 장 유스 케이스 분석

제 1 절 유스 케이스 지침 설계

본 연구가 제안하는 시스템이 다양한 환경에 적용 가능한지 확인하기 위하여 다음 조건을 충족시키는 데이터 세트과 모델을 사용한 프로토타입을 제작했다. 첫째, 수치 예측 모델과 분류 예측 모델을 모두 실험한다. 둘째, 분석 도구의 도움 없이는 해석이 어려운 정도로 모델의 구조를 충분히 복잡하게 한다. 셋째, 수치형 데이터와 범주형 데이터가

복합적으로 구성된 데이터셋을 사용한다. 이에 Insurance Premium Prediction 데이터셋과 이를 학습한 인공 신경망 모델을 활용하고자 한다³⁴. Insurance Premium Prediction 데이터셋은 나이, 성별, BMI, 흡연 여부, 거주지역, 의료 비용 지출 등 수치 및 범주로 표현 가능한 입력 변수를 포함한다. 본 유스 케이스에서는 의료 비용 지출액을 나머지 입력 변수를 통해 예측하는 인공 신경망 모델을 개발하여 사용했다.

이 사례 연구에서는 시스템의 사용 방법과 이점에 대한 전반적인 설명을 제공한다. 이 절에서는 설계 목표(G1 ~ G4)에 대하여 질문을 세우고, 그 질문에 답을 얻는 과정을 설명함으로써 본 연구가 제안하는 시스템의 사용성을 설명한다.

- Q1: 머신러닝 모델의 예측 값을 높게 만드는 입력 변수 목록은 무엇인가?
- Q2: 머신러닝 모델의 예측 값과 데이터의 실제 값이 크게 다른 인스턴스들은 공통적으로 어떤 입력 변수 값 분포를 가지고 있는가?
- Q3: Q1에서 선정된 변수 중 하나가 변경될 경우 머신러닝 모델이 예측한 의료 비용 지출 예측 값 변화는 어떠한가?
- Q4: Q1에서 선정된 변수 중 하나의 값이 큰 그룹에서, Q1에서 선정된 변수 중 다른 하나가 변경될 경우 흡연 여부 변경이 모델의 예측 값 미치는 영향을 보여주는 개별 인스턴스에는 무엇이 있는가?

³⁴ Kaggle, Insurance Premium Prediction,
Kaggle(<https://www.kaggle.com/noordeen/insurance-premium-prediction>), 2019.

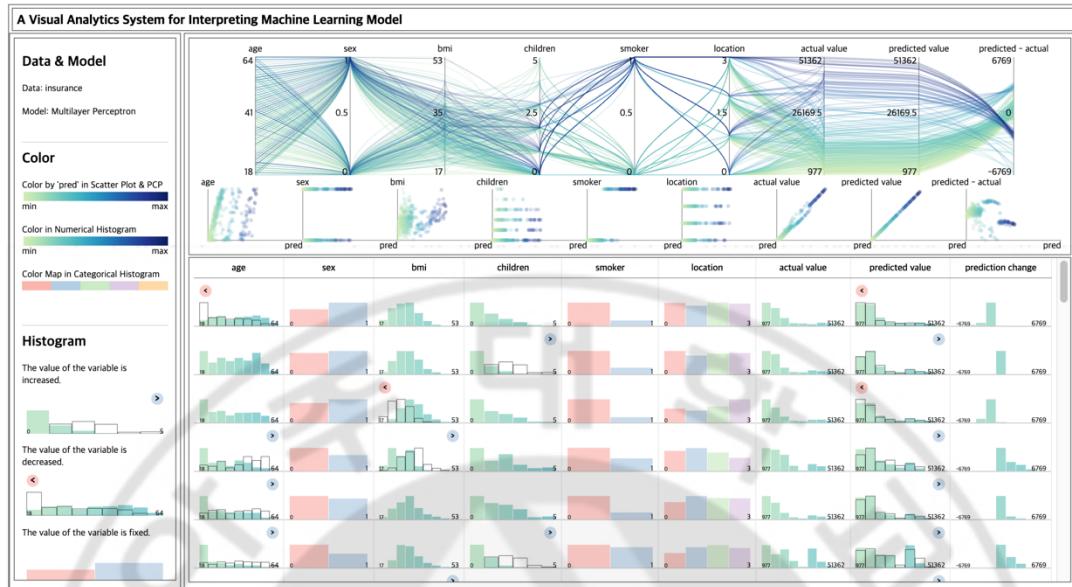


그림 19. 유스 케이스에 해당하는 데이터를 시각화한 모습. Insurance Premium

Prediction 데이터셋과 이를 학습한 모델을 해석한다.

제 2 절 유스 케이스 분석 수행

Q1: 머신러닝 모델의 예측 값을 높게 만드는 입력 변수 목록은 무엇인가?

이 질문은 평행 좌표 그래프의 predicted value 축의 값이 높은 인스턴스를 필터링하여 확인할 수 있다. Predicted value 축은 타깃 변수의 모델 예측 값을 표시한다. 따라서, 이 값이 높은 인스턴스를 추려내어, 이들의 변수 분포를 살펴보는 것이 질문에 대한 대답을 내는 데 도움이 된다. 사용자는 필터링된 인스턴스의 입력 변수 값 분포를 산점도에서 살펴볼 수 있다. 다른 변수에서는 큰 경향성을 찾기 어렵지만, smoker(흡연 여부) 변수에서는 차이가 있다. 의료 비용 지출 예측 값이 높은 대부분의 인스턴스가 흡연자로 기록되어 있음이 산점도에 나타난다. 또한 흡연 여부보다는 미미한 경향성이지만 BMI 지수와 의료 비용 지출 예측 값에도 양의 상관관계가 있음이 보인다.

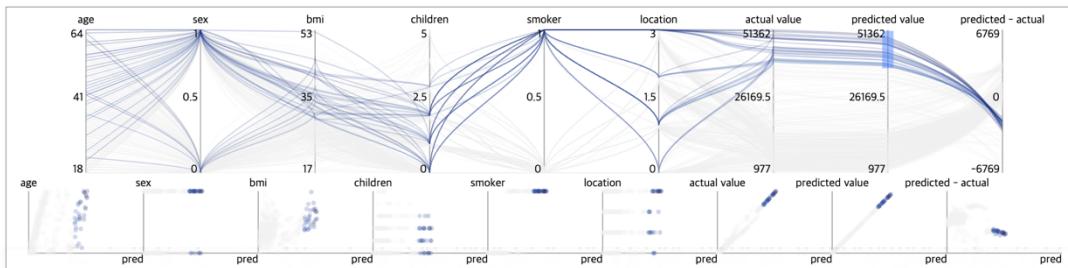


그림 20. 타깃 변수의 모델 예측 값이 높은 인스턴스를 필터링한 모습. 평행 좌표 그래프에서 마우스 드래그 인터랙션을 사용해 인스턴스를 필터링할 수 있다. 이 그룹은 대체로 높은 BMI 지수를 가지며 흡연을 하는(smoker=1) 인스턴스로 구성되어 있다.

Q2: 머신러닝 모델의 예측 값과 데이터의 실제 값이 크게 다른 인스턴스들은 공통적으로 어떤 입력 변수 값 분포를 가지고 있는가?

이 질문 역시 평행 좌표 그래프 필터링과 산점도 시각화 분석을 통해 확인할 수 있다. 사용자는 이 질문에 대답하기 위해서, 실제 값과 모델 예측 값의 차이를 표시하는 축을 통해 인스턴스를 필터링할 수 있다. 예측 값과 실제 값이 다른 경우는 예측 값이 큰 경우와 작은 경우 둘로 나눌 수 있다. 이번 사례에서는 예측 값이 실제 값보다 큰 사례를 살펴본다. 평행 좌표 그래프와 산점도의 분포를 통해 흡연 여부가 참인 인스턴스에서 실제 값보다 모델 예측 값이 크게 예측됨을 확인할 수 있다. 여기서 독특한 점은 대부분의 인스턴스가 실제 의료 비용 지출 값이 중간 값 정도에 분포하고 있다는 점이다. 이를 통해 사용자는 ‘모델이 중간 값 정도의 의료 비용을 지출하는 사람이 흡연할 경우 실제보다 크게 의료 비용을 예측한다.’고 추론할 수 있다.

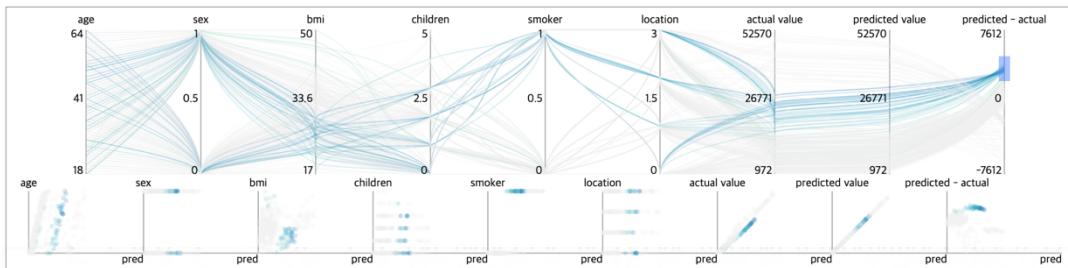


그림 21. 타깃 변수의 모델 예측 값이 실제 값보다 큰 인스턴스를 필터링한 모습. 이 그룹은 대체로 중간 값 정도의 실제 보험료를 가진 흡연자로 구성되어 있다.

Q3: Q1에서 선정된 변수 중 하나가 변경될 경우 머신러닝 모델이 예측한 의료 비용 지출 예측 값 변화는 어떠한가?

이 질문은 Q1, Q2와 다르게 시스템 하단 테이블 영역을 함께 활용하여 분석해야 한다. 이 설명에서는 Q1에서 선정된 변수 중 하나를 `smoker`로 선택하였다. 테이블 영역이 변수가 변형된 인스턴스의 값을 보여주기 때문이다. 테이블 영역에서 `smoker` 열의 이름을 클릭해, 흡연 여부가 참에서 거짓 또는 거짓에서 참으로 변경된 그룹을 찾을 수 있다. 이들의 `predicted value` 열(모델 예측 값)을 확인하면 흡연 여부의 변경이 모델 예측에 미치는 영향을 확인할 수 있다. 그림 22 와 그림 23 은 흡연 여부를 거짓에서 참으로 바꿀 경우 예측된 비용이 증가하고, 반대의 경우에는 감소하는 것을 보여준다. 사용자는 이를 통해 ‘흡연 여부가 모델의 의료 비용 예측에 영향을 주며, 그 방향성은 양적(positive)이다.’라고 결론지을 수 있다. 이들은 다른 변수가 고정된 상황에서의 흡연 여부의 한계 효과(marginal effect)를 살펴본 것이기 때문에 평행 좌표 그래프와 산점도에서 확인한 것보다 더 확신할 수 있는 정보이다.

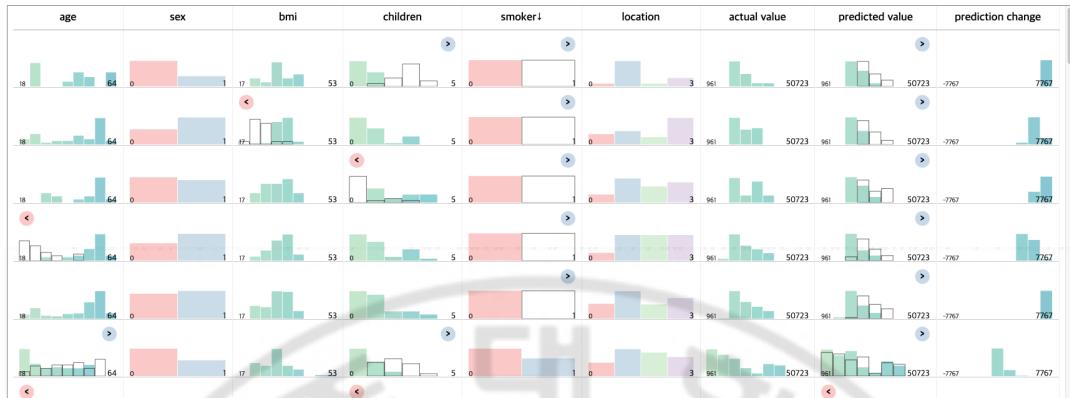


그림 22. 흡연 여부가 거짓에서 참으로 변경된 그룹. 표에서 행의 이름을 한 번 클릭하여 맨 위로 정렬시켰다. 테이블의 각 칸에 표시된 화살표의 방향과 그를 둘러싼 색칠된 동그라미, 그리고 히스토그램이 값의 변화를 표시한다. 원본 데이터에서는 색칠된 히스토그램의 분포를 가지며, 변경된 데이터에서는 스트로크로만 표시된 히스토그램의 분포를 가진다.

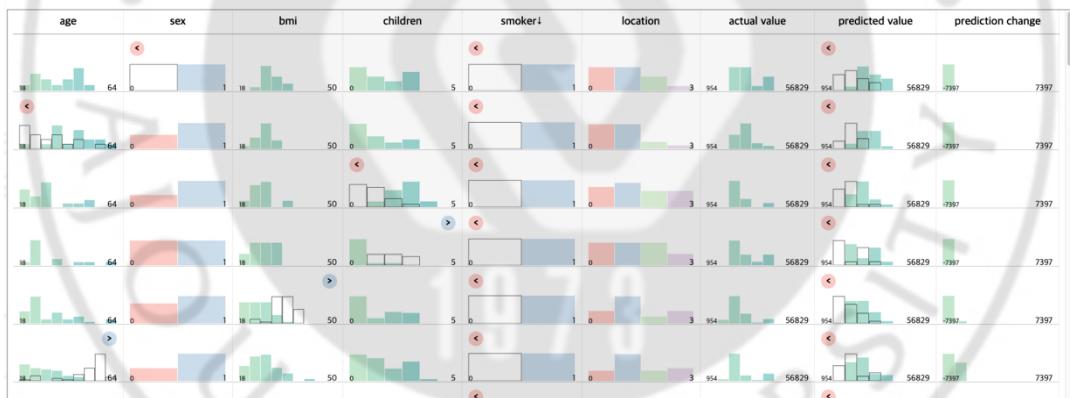


그림 23. 흡연 여부가 참에서 거짓으로 변경된 그룹. 표에서 행의 이름을 두 번 클릭하여 맨 위로 정렬시켰다.

Q4: Q1에서 선정된 변수 중 하나의 값이 큰 그룹에서, Q1에서 선정된 변수 중 다른 하나가 변경될 경우 흡연 여부 변경이 모델의 예측 값 미치는 영향을 보여주는 개별 인스턴스에는 무엇이 있는가?

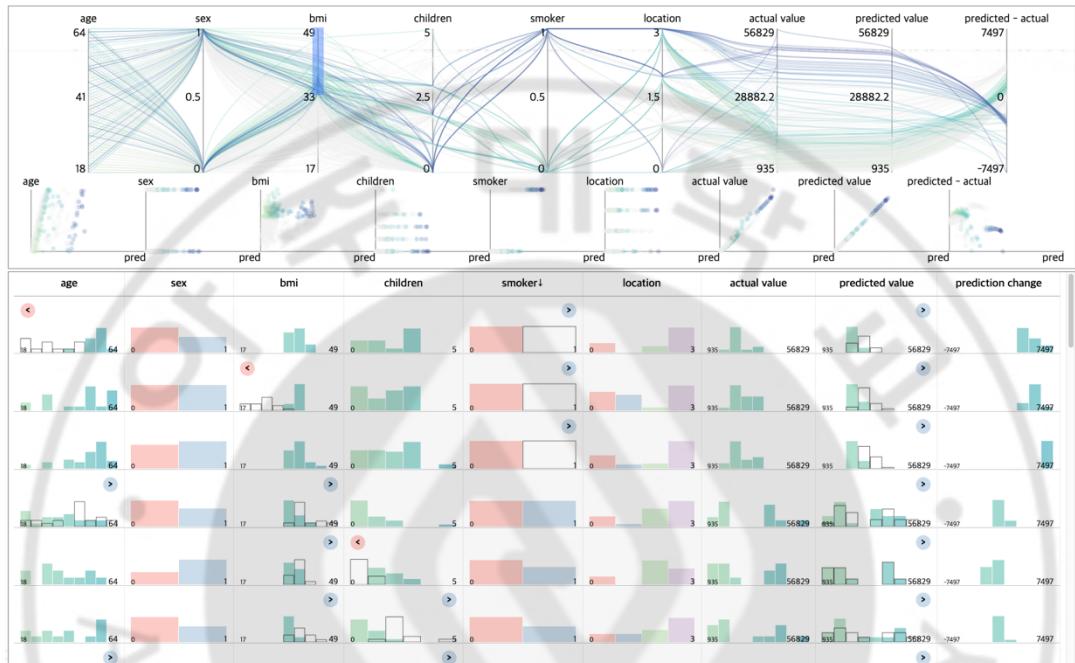


그림 24. BMI 가 높은 그룹을 평행 좌표 그래프에서 필터링하고, 이들의 흡연 여부가 음에서 양으로 바꾼 그룹을 테이블 맨 위로 정렬시킨 모습. 표에서 맨 위의 그룹은 나이와 흡연 여부가 변경된 경우를 표시한다. 이 그룹을 선택할 경우, 자세히 보기에서 나이가 줄어들고 비흡연에서 흡연으로 변경되었을 때의 모델 예측 변화 추이를 볼 수 있다. 위에서 세 번째 그룹은 흡연 여부만 변경되었다. 이 그룹을 선택할 경우 자세히 보기에서 흡연 여부의 의료 비용 예측 값에 대한 한계 효과를 살펴볼 수 있다.

이 질문의 목적은 한 변수(A 변수)의 값을 특정 범위에 고정한 상태에서, 다른 변수(B 변수)의 값이 변화할 때 모델의 예측이 어떻게 바뀌는지를 살펴보는 작업을 진행하도록 하는 것이다. 이 사례에서는 A 변수로 BMI, B 변수로 smoker 를 선정하였다. 이 질문에

답하기 위해서는 다음의 과정이 필요하다. 첫째, BMI 가 높은 그룹을 평행 좌표 그래프에서 필터링한다. 이는 사용자의 마우스 드래그 인터랙션으로 가능하다. 둘째, 표에서 smoker 열을 클릭해 흡연 여부가 변경된 인스턴스를 맨 위로 정렬한다. 셋째, 정렬된 그룹 중 흡연 여부만 변경된 행(그림 24 속 테이블 시각화의 세 번째 행)을 마우스로 클릭하면 시스템은 이 조건을 갖춘 인스턴스 목록을 상세 보기로 띠워준다. 이곳에서 테이블 정렬 기능을 통해 변화 값이 크거나 작은 인스턴스를 대표 인스턴스로 선정할 수 있다. 입력 변수 변화 값이 큰 인스턴스를 선정할 경우, 예측 값 변화의 폭을 보여주는 데 유리하다. 반면, 입력 변수 변화 값이 작은 인스턴스를 선정할 경우에는, 입력 변수의 작은 변화가 큰 예측 변화로 이어지지는 않는지에 대한 안정성 검증을 수행할 수 있다. 대표 인스턴스는 사용자의 상황에 따라 다른 속성을 가지므로 본 시스템의 정렬 인터랙션이 유용할 수 있다.

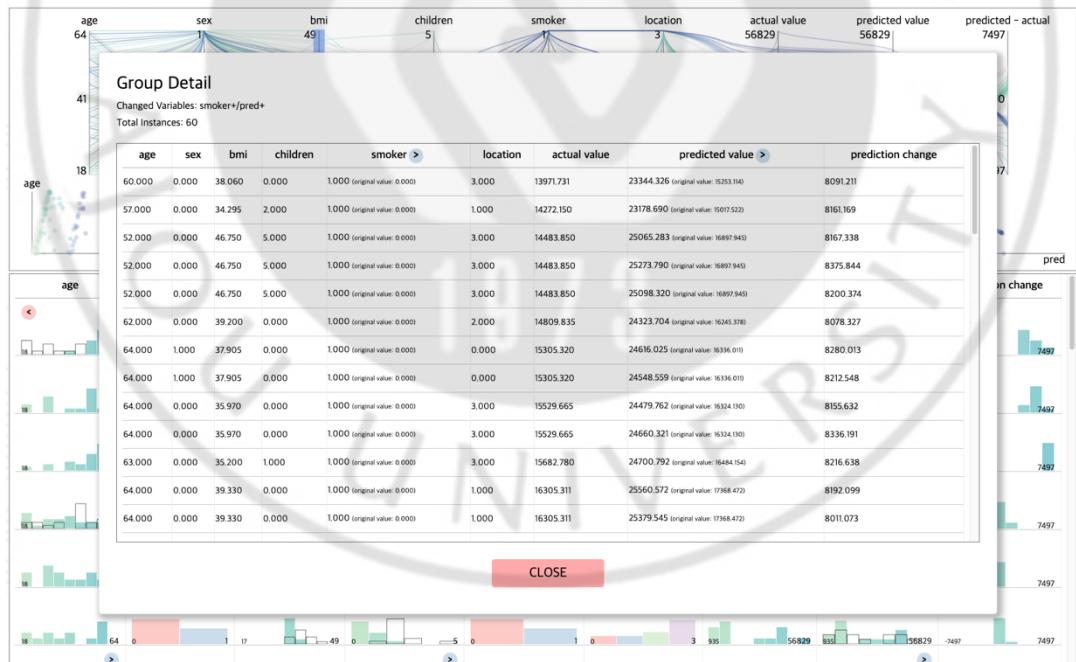


그림 25. BMI 가 높은 인스턴스의 흡연 여부를 음에서 양으로 바꾼 그룹의 상세 보기

제 6 장 사용자 심층 인터뷰

제 1 절 사용자 심층 인터뷰 설계

본 연구에서는 입력 변수가 변형된 인스턴스 생성과 이를 탐색하는 필터링, 그룹, 정렬 등의 다양한 인터랙션이 사용자의 모델 해석에 도움이 된다는 연구질문을 바탕으로 시각 분석 시스템을 설계 및 개발하였다. 제안된 시스템의 사용성 검증을 위하여, 3 명의 대상자를 바탕으로 심층 인터뷰를 진행하였다. 안재욱의 연구는 특정한 데이터를 분석하는 도구와 기법을 평가하기 위한 지표를 제안했는데 이는 포괄적 이해(Comprehensiveness), 탐색 가능성(Discoverability), 유용성(usefulness), 용이한 정도(Easiness) 등을 포함한다³⁵. 이 평가 지표들은 머신러닝 모델의 해석을 위한 시각화 도구에도 적용될 수 있다.

1) 포괄적 이해: 제안된 도구와 기법을 활용하여 데이터를 전반적으로 이해하는 데 도움이 되는지 평가한다. 머신러닝 모델 해석의 경우에는 제안된 도구가 머신러닝 데이터 셋의 개요와 및 모델의 전반적인 결정 패턴을 파악하는 데 도움이 되는지 평가할 수 있다.

2) 탐색 가능성: 제안된 도구와 기법을 통해 목표를 달성하기 위해 필요한 업무가 무엇인지 탐색하는 데 도움이 되는지 평가한다. 머신러닝 모델 해석의 경우에는 사용자가 모델 해석을 위해 어떤 입력 변수, 인스턴스, 인스턴스 그룹 등을 살펴봐야 하는지 아는 데 제안된 도구가 도움이 되는지 평가할 수 있다.

³⁵ Ahn, Jae-wook, Catherine Plaisant, and Ben Shneiderman. A task taxonomy for network evolution analysis. IEEE transactions on visualization and computer graphics 20.3 2013.

3) 유용성: 제안된 도구를 통해 구체적인 분석 업무 및 질문에 대해 명확한 분석 결과를 얻을 수 있는지 평가한다. 머신러닝 해석을 위한 도구의 경우에는 위 (탐색 가능성) 부분에서 도출된 질문에 명확한 답을 찾아낼 수 있는지 평가할 수 있다.

4) 용이한 정도: 제안된 도구 및 기법을 분석에 활용하는 방법을 이해하기 쉽거나 어려운 정도를 평가한다. 머신러닝 해석의 경우에도 제안된 도구를 타깃 사용자가 이해하고 활용하는 데 어려움이 있는지 조사할 수 있다.

본 연구자는 인터뷰를 통해 제안된 시스템의 전반적인 사용 난이도 및 유용성을 파악하였고, 나아가 개별 기능의 존재 여부가 시스템 사용성에 미치는 영향을 파악하였다. 뿐만 아니라, 인터뷰를 통하여 제공된 시스템이 설계 목표(G1 ~ G4)를 달성하기 위하여 데이터의 다양한 측면을 탐색하는 데 도움이 되는지 살펴보았다.

인터뷰 대상자는 전문가 1명과 비 전문가 2명으로, 비 전문가는 각각 실무 경험이 없는 전공자와 간단한 실무 경험이 있는 비전공자로 구성하였다. 전문가 인터뷰는 시스템의 기여점과 한계를 파악하는 데 주요 초점을 맞췄고, 비 전문가 인터뷰는 타깃 사용자가 시스템을 이해하고 활용하는 데에 어려움과 혼동이 있는지 파악하는 데 초점을 맞췄다. 각각 인터뷰 대상자의 특성과 주요 인터뷰 항목은 다음과 같다.

1) 데이터 분석 전문가(P1): 이 대상자는 데이터 분석 분야 석사학위 이상에 준하는 학력을 가진 자로 실무 환경에서 데이터를 가공, 분석, 시각화하는 과정을 설계해본 경험이 다수 있다. 이 대상자는 머신러닝 및 데이터 분석 분야의 이론적 지식을 알고 있으며, 실무 경험 또한 풍부하다. 이 대상자를 인터뷰하는 목적은 본 연구가 제안하는 시스템의 개별 기능이 실효성을 가지는지 조사하기 위함이다. 이 대상자와의 인터뷰로부터 입력 변수 변형 데이터 사용 여부가 모델 해석을 용이하게 만드는지

확인하였다. 또한, 필터, 그룹, 정렬 등 인터랙션 적용 여부가 시스템 사용성 재고에 미치는 영향을 분석했다.

2) 데이터 분석 실무 경험이 없는 전공자(P3): 이 대상자는 데이터 처리 및 머신러닝 등 데이터 분석의 이론적 측면을 공부한 경험이 있는 컴퓨터 공학 학사 이상의 학력을 지닌 자로 실무 환경에서 데이터 분석을 경험한 적은 없고, 데이터 시각 분석에 대한 경험도 없다. 이 대상자를 인터뷰하는 목적은 실무 경험이 부족한 사용자가 본 연구가 제안하는 시스템을 쉽게 이해하고 사용하는지 파악하기 위함이다. 이 대상자와의 인터뷰를 통하여 본 연구가 제안하는 시스템의 활용하는 난이도와 분석 유용성을 파악했다.

3) 데이터 분석 실무 경험이 있는 비전공자(P3): 이 대상자는 일반적인 업무 수준에서 데이터를 분석해본 경험이 있는 자로 공학을 전공하지 않았다. 또한, 데이터 시각 분석에 대한 경험도 없었다. 엑셀 및 스프레드 시트 등을 활용하여 기초 통계 작성 및 조회 수준에서 데이터 분석 업무를 수행해 왔다. 이 대상자는 머신러닝 모델이 기존 데이터를 학습하여 새로운 상황의 결과를 예측하는 블랙박스라는 내용 수준의 사전 지식을 가지고 있다. 이 대상자를 인터뷰하는 목적은 제안된 시스템이 머신러닝 지식이 부족한 사용자가 머신러닝 모델 해석에 도움이 되는지 확인하는 데 있다. 이 대상자와의 인터뷰를 통하여 본 연구가 제안하는 시스템의 활용 난이도와 분석 유용성을 파악했다.

심층 인터뷰 진행은 각 대상자 별로 약 1 시간 30 분이 소요되었다. 인터뷰는 설명 세션 45분과 사용 및 평가 세션 45분으로 나뉜다. 설명 세션에서는 연구 개요와 목적, 같은 목적을 공유하는 기존 연구 사례 그리고 설계된 시스템을 소개했다. 이때, 유스 케이스 분석에 활용된 의료 비용 예측 모델 분석 프로토타입을 바탕으로 시스템을 사용했다. 사용 및 평가 세션에서는 대상자 스스로 설계 목표(G1 ~ G4)와 관련된

질문을 세우고 이에 대해 스스로 탐색하도록 하였다. 이때, 당뇨 여부 분류 모델 분석 프로토타입을 통해 사용자가 모델을 분석해보게 하면서 시스템 전체와 개별 기능 모두에서 활용 난이도와 분석 유용성을 평가하도록 했다. 특히, 이 과정에서 본 연구의 주요 제안 사항인 실제 데이터와 변형된 데이터 모두를 활용하여 모델의 예측 패턴을 해석하는 일을 수행하는 데 본 연구가 제안하는 시스템의 효과를 살펴보았다.

사용자 인터뷰의 사용 및 평가 세션에서는 Pima Indian Diabetes 데이터 셋과 이를 학습한 로지스틱 회귀 모델을 분석하는 사례를 사용하였다³⁶. Pima Indian Diabetes 데이터 셋은 임신 횟수, 포도당 부하 검사 수치, 혈압, 혈청 인슐린 등의 의료 정보와 함께 대상자의 당뇨병 유무를 기록한 데이터이다. 본 인터뷰 세션에서는 이 데이터를 활용하여 의료 정보를 토대로 당뇨병 여부를 예측하는 로지스틱 회귀 모델을 개발하여 활용하였다.

³⁶ Kaggle, Pima Indian Diabetes, Kaggle(<https://www.kaggle.com/uciml/pima-indians-diabetes-database/metadata>), 2016.

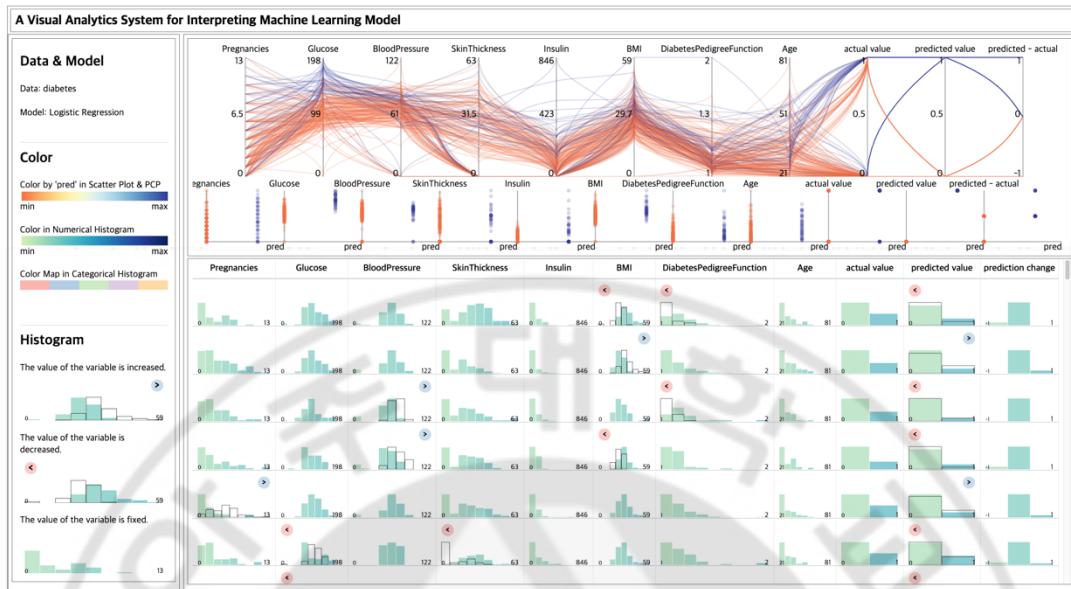


그림 26. 제안된 시스템에 Pima Indian Diabetes 데이터 세트과 이를 학습한 로지스틱 회귀 모델이 적용된 모습

제 2 절 사용자 심층 인터뷰 수행 및 결과

사용자 인터뷰에서 모든 사용자는 본 연구가 제안하는 시스템을 이해하고 활용하여 머신러닝 모델을 분석하는 데 큰 어려움을 겪지 않았다. 인터뷰로부터 다음 세 가지 발견점을 얻을 수 있었다. 첫째, 본 연구가 제안하는 시스템은 사용자가 실제 현상이 아니라 모델의 작동 패턴을 해석하는 것에 초점을 맞춰 분석 작업을 하도록 유도하는 데 도움을 주었다. 사용자는 제안된 시스템을 통해 데이터 및 모델을 포괄적으로 이해하고 분석에 필요한 업무 목록을 수월하게 파악했다. 둘째, 본 연구가 제안하는 시스템은 머신러닝 비 전문가 사용자층으로 하여금, 실제 데이터와 변형된 데이터를 각각 분석 목표에 맞춰 적절하게 사용하도록 도왔다. 사용자는 제안된 시스템을 활용해 원하는 분석 작업에 대한 명확한 답을 구할 수 있었다. 셋째, 본 연구가 제안한 시스템의 시각화

요소 사용 및 레이아웃 구성은 머신러닝 및 데이터 시각 분석에 익숙하지 않은 사용자에게도 이해와 활용이 어렵지 않았다.

머신러닝 및 데이터 분석 분야의 전문 지식이 부족한 P2는 설명 세션에서 해석 가능한 머신러닝 분야가 ‘실제 현상을 해석하는 것이 아니라 머신러닝 모델의 예측 결과를 해석’하는 목적을 가졌다는 점을 이해했다. 하지만, 기존 연구 사례인 Partial Dependence Plot의 결과물을 해석함에 있어서 실제 현상과 모델 예측이 차이를 구분짓는 데 혼란을 겪었다²⁰. 예를 들어, 온도가 높을수록 자전거 대여 예측 값이 올라가는 결과를 P2는 온도가 높을수록 자전거 대여 실제 값이 올라가는 것으로 오인했다. 반면, 해당 대상자는 본 연구의 결과물을 통해 모델을 해석할 때에는 실제 값과 모델 예측 값을 쉽게 구분했다.

P2는 평행 좌표 그래프에 실제 값을 표시하는 축(actual value)과 예측 값을 표시하는 축(predicted value)가 각각 존재하는 것이 이 둘을 구분하는 데 도움이 되었다고 평가했다. 또한 P3는 평행 좌표 그래프에 예측 값과 실제 값의 차이를 표시하는 축(predicted – actual)이 있어서, 모델의 정확도를 높이는 입력 변수와 오류를 크게 만드는 입력 변수를 파악하는 데 도움이 되었다고 말했다. P3는 예측 값과 실제 값의 차이를 표시하는 축에서 예측 오류가 큰 인스턴스를 마우스 드래그 인터랙션으로 필터링하고, 산점도를 확인하여 필터링된 인스턴스가 다른 입력 변수에서 어떤 분포를 보이는지 파악하였다. 이 점을 미루어보아, 머신러닝 실무 경험이 부족한 사용자에게 평행 좌표 그래프의 축 활용이 모델 해석에 도움을 주는 점을 알 수 있었다. 이들 축의 활용을 통해 인터뷰 대상자들은 모델 해석을 위해 먼저 살펴봐야 할 인스턴스 목록을 추론했다. 뿐만 아니라, P3는 데이터 셋과 모델에 대한 설명이 없는 상황에서도 필터링 및 정렬을 통한 모델 특성 파악 업무를 수월하게 진행하였다. 평행 좌표 그래프와

산점도를 활용하여 극한값을 가지거나 특정 분포를 띠는 인스턴스를 필터링하는 등의 분석 작업을 수행했다. P3는 인터뷰에서 ‘시각적으로 튜는 요소가 궁금증을 유발했고, 이를 자연스럽게 먼저 살펴보게 되었다’고 설명했다. 이러한 인터뷰 결과를 통해 평행 좌표 그래프 및 산점도의 활용이 사용자로 하여금 필요한 업무가 무엇인지 탐색하는 데 긍정적인 영향을 주었다고 평가할 수 있다.

모든 인터뷰 대상자는 제안된 시스템이 실제 데이터와 변형된 데이터 모두를 활용하여 모델을 분석한다는 점을 이해하고, 이를 분석 업무에 활용했다. P2는 평행 좌표 그래프 및 산점도가 표시된 시스템 상단 영역에 실제 데이터가 표시되고, 테이블 형태로 히스토그램이 배치된 시스템 하단 영역에 변형된 데이터가 표시되는 점이 두 가지 데이터를 구분할 수 있는 요인이라고 언급했다. 실제 데이터와 변형된 데이터가 각각 물리적으로 구분된 공간에 표시되기 때문에 이를 구분 짓기 편하다는 의미이다.

P3는 레이아웃에 더해서 시각적 표현의 장점이 이 두 가지 데이터를 구분하게 만드는 데 도움이 된다고 평가했다. 색칠된 히스토그램(실제 값)과 스트로크로만 표현된 히스토그램(변형된 값)을 동시에 시각화하고, 이를 변형된 값의 방향성에 따라 파란색(값 커짐) 또는 빨간색(값 작아짐) 바탕을 가진 화살표와 함께 표현한 점을 장점으로 언급했다. 기존 연구에서는 변형된 값과 기존 값의 색상, 위치, 모양 등이 다르지 않아서 분석에 어려움이 있었던 한계점을 극복한 것이라고 볼 수 있다. 본 연구에서 화살표 및 색칠 여부가 다른 히스토그램 기법을 도입할 수 있었던 이유는 변형된 인스턴스를 변형된 변수의 종류 및 변화 방향(상승 및 하락 등)에 따라 그룹 지었기 때문이다. 위와 같은 인터뷰 내용을 통해 그룹 기능과 레이아웃 배치와 시각적 요소 사용이 시스템의 용이한 정도를 높이는 데 긍정적 효과를 줬다고 해석할 수 있다.

또한, 본 인터뷰를 통해 시스템 레이아웃 및 구성이 적절하게 되었음을 확인했다. P2 와 P3 는 별도의 지시 없이도 평행 좌표 그래프 확인, 산점도 확인, 필터링 인터랙션, 표, 정렬 인터랙션, 그룹 상세 보기 팝업 창 확인의 순서로 모델 해석 작업을 진행했다. 이 순서는 4 장에서 소개한 분석 시스템이 작동하는 구조와 일치한다. 사용자가 이들 기능을 순차적으로 사용하도록 유도할 수 있었던 이유는 이들 기능이 페이지 상단에서 하단으로 순차적으로 배치되어 있을 뿐만 아니라, 각각의 기능이 시각화 만트라가 제안하는 오버 뷰(평행 좌표 그래프 및 산점도 확인), 필터링 및 정렬(평행 좌표 그래프 필터링 및 테이블 정렬), 자세히 보기(그룹 상세 보기 팝업)로 적절히 구분되어있기 때문이라고 볼 수 있다²⁹. 본 연구가 제안한 방식의 분석 시스템 구축이 사용자가 데이터와 머신러닝 모델을 포괄적으로 이해하고 필요한 업무가 무엇인지 탐색하는 데 도움을 주었다고 평가할 수 있다.

인터뷰를 통해 발견된 본 연구의 한계점 및 개선 요구 사항으로는 다음과 같은 사항이 있었다. 우선, 시각적 요소 해석의 어려움이 있었다. 평행 좌표 그래프에서 인스턴스를 표시하는 선이 꼬여 있거나 겹쳐 있는 경우가 다수 있어서 어떤 선이 어떤 선을 의미하는지 파악하기 어렵다는 피드백이 P1 과 P3 로부터 나왔다. 특히 인터뷰 P1 은 다른 시각 분석 시스템의 사례를 들어 시스템 사용성을 높이기 위해 축 순서의 변경, 축에 인스턴스의 빈도수 표시 등의 기능을 필요하다고 평가했다³⁷. 다음으로 구체적인 값 파악이 어렵다는 점이 아쉬운 점으로 꼽혔다. P1 은 데이터의 개요를 보다 정확하게 파악하기 위해 입력 변수간 상관계수를 보여주거나 인스턴스 그룹의 입력 변수 별 평균

³⁷ Steed, C. A., Shipman, G., Thornton, P., Ricciuto, D., Erickson, D., & Branstetter, M., Practical application of parallel coordinates for climate model analysis. Procedia Computer Science 9, 2012.

및 분산 등 통계 정보가 필요하다고 지적했다. 본 연구자는 이와 같은 피드백이 합당할 뿐만 아니라, 제안된 시스템에 쉽게 통합될 수 있는 기능으로 파악하였다. 이에 따라, 평행 좌표 그래프의 가시성 향상 및 통계치 정보 표시를 향후 개선 요구 사항으로 지정하였다.

제 7 장 결론

본 연구는 데이터의 입력 변수가 모델의 예측 결과에 어떤 영향을 주는지를 바탕으로 머신러닝 모델을 해석하는 시각 분석 시스템을 제안했다. 이를 위하여 SMOTE 를 활용해 데이터 입력 변수 값을 변형하는 기술을 제안하고 변형된 데이터에 대한 모델 예측 값 변화 변화 추이를 시스템이 추적하도록 했다. 이 방식은 입력 데이터 값을 합리적인 범위 내에서 변화시킨다는 점에서 기존 연구의 한계를 극복했다. 또한, 제안된 시스템은 시각화 인터페이스를 통해 사용자의 모델 해석을 용이하게 했다. 머신러닝 예측 결과에 대해 필터, 그룹, 정렬 인터랙션을 적용했으며, 평행 좌표 그래프와 산점도, 테이블 시각화 등 다양한 시각화 양식을 조합했다. 이러한 기법이 모델 해석에 긍정적인 영향을 미침을 유스 케이스와 사용자 인터뷰를 통해 확인하였다. 본 연구의 기여는 머신러닝 비 전문가가 머신러닝 모델 해석 기법을 이해하고, 머신러닝 모델을 해석하는데 도움이 되는 시각 분석 시스템을 설계하고 평가했음에 있다.

본 연구에서는 머신러닝 모델을 해석하도록 지원하는 시각화 기반 분석 시스템을 제안하였다. 이는 머신러닝 모델과 사람이 상호 협력하도록 돋는 시스템이다. 머신러닝 모델이 자동적으로 분류와 예측 작업을 수행하도록 하고, 사용자는 모델의 결과를 도입하기 전에 분석 시스템을 통해 모델이 학습한 바가 무엇인지 파악하고, 그 적절성을 확인한다. 이 과정을 통해 사용자는 모델과 데이터 셋을 작동 패턴을 이해하고, 한계를

인식하며, 부작용에 대응할 수 있게 된다. 심층 사용자 인터뷰에서 머신러닝 비 전문가인 사용자들은 본 연구가 제안하는 분석 도구를 통해 머신러닝의 개념과 해석 기법의 작동 방식을 이해하는 데 도움을 받았다고 평가했고, 실제 모델 분석 과정에서 유의미한 분석 결과를 도출해낼 수 있었다.

이 연구의 한계점은 다음과 같다. 우선, 인터뷰에서 밝혀진 바와 같이, 평행 좌표 그래프의 가시성 향상 및 통계치 정보 표시를 통해 시스템의 유용성 및 용이한 정도를 높일 필요가 있다. 또한, 시스템의 효과 검증 측면에서 대규모 데이터를 바탕으로 한 확장성 검증이 부족했다는 한계점을 갖는다. 향후 연구에서는 이러한 한계점을 극복하기 위하여 정확한 수치 파악을 위한 통계치 표시 기능과 평행 좌표 그래프 및 산점도의 가시성을 높이는 방안을 도입할 필요성이 있다. 본 연구가 제안하는 시스템의 주요 목적은 입력 변수가 모델 출력에 미치는 영향을 판단하는 것이다. 이 주제를 다루는 최근의 연구에서 머신러닝 기법을 활용하여 개별 입력 변수가 모델 출력에 미치는 영향을 수치화하는 변수 중요도(Feature Importance) 연구가 진행되고 있다³⁸. 향후 연구에서 이러한 변수 중요도를 함께 시각화하면 사용자는 보다 정확하게 주요 변수를 파악할 수 있을 것이다.

³⁸ Li, X., et al. A debiased MDI feature importance measure for random forests. arXiv 2019.

참고문헌

1. Molnar, Christoph. Interpretable machine learning. Lulu.com, 2020.
2. Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. Explaining explanations: An overview of interpretability of machine learning. 2018 IEEE 5th International Conference on data science and advanced analytics (DSAA). IEEE, 2018.
3. Vilakone, P., Park, D. S., Xinchang, K., & Hao, F. An efficient movie recommendation algorithm based on improved k-clique. Human-centric Computing and Information Sciences 8.1: 1–15, 2018.
4. Tanveer, Muhammad Suhaib, Muhammad Umar Karim Khan, and Chong-Min Kyung. Fine-tuning darts for image classification. 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021.
5. Feifel, Patrick, Frank Bonarens, and Frank Koster. Reevaluating the Safety Impact of Inherent Interpretability on Deep Neural Networks for Pedestrian Detection. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.
6. Das, D., Ito, J., Kadowaki, T., & Tsuda, K. An interpretable machine learning model for diagnosis of Alzheimer's disease. PeerJ 7: e6543, 2019.
7. Kurzweil, R. The singularity is near: When humans transcend biology; Penguin, 2005.
8. Thiel, P.A.; Masters, B. Zero to one: Notes on startups, or how to build the future; Currency, 2014.

9. Doshi-Velez, Finale, and Been Kim. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608, 2017.
10. Van Den Elzen, S., & Van Wijk, J. J., Van Den Elzen, Stef, and Jarke J. Van Wijk. Baobabview: Interactive construction and analysis of decision trees. 2011 IEEE conference on visual analytics science and technology (VAST). IEEE, 2011.
11. Mitchell, Michael N. Interpreting and visualizing regression models using Stata. Vol. 558. College Station, TX: Stata Press, 2012.
12. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). International conference on machine learning. PMLR, 2018.
13. Wang, Z. J., Turko, R., Shaikh, O., Park, H., Das, N., Hohman, F., ... & Chau, D. H. P. CNN explainer: Learning convolutional neural networks with interactive visualization. IEEE Transactions on Visualization and Computer Graphics 27.2, 2020.
14. Ren, D., Amershi, S., Lee, B., Suh, J., & Williams, J. D.. Squares: Supporting interactive performance analysis for multiclass classifiers. IEEE transactions on visualization and computer graphics 23.1, 2016.
15. Molnar, Christoph, Giuseppe Casalicchio, and Bernd Bischl. Interpretable machine learning – a brief history, state-of-the-art and challenges. Joint

European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, Cham, 2020.

16. Park, H., Nam, Y., Kim, J. H., & Choo, J., HyperTendril: Visual Analytics for User-Driven Hyperparameter Optimization of Deep Neural Networks, *IEEE Transactions on Visualization and Computer Graphics*, 27(2), 2020.
17. Chanhee Park, Hyojin Kim, and Kyungwon Lee. A Visualization System for Performance Analysis of Image Classification Models. *Electronic Imaging* 2020.1, 2020.
18. Zhang, J., Wang, Y., Molino, P., Li, L., & Ebert, D. S. Manifold: A model-agnostic framework for interpretation and diagnosis of machine learning models. *IEEE transactions on visualization and computer graphics* 25.1, 2018.
19. Park, Y., and J. Y. Yun. A Design Case Study of Artificial Intelligence Pipeline Visualization. *Archives of Design*, 2021.
20. Ming, Yao, et al. Understanding Hidden Memories of Recurrent Neural Networks, 2017 IEEE Conference on Visual Analytics Science and Technology (VAST). IEEE, 2017.
21. Kahng, M., Andrews, P. Y., Kalro, A., & Chau, D. H., ActiVis: Visual Exploration of Industry-Scale Deep Neural Network Models, *IEEE transactions on visualization and computer graphics* 24.1, 2017.
22. Ming, Yao, Huamin Qu, and Enrico Bertini. Rulematrix: Visualizing and understanding classifiers with rules. *IEEE transactions on visualization and computer graphics* 25.1, 2018.

23. Smilkov, D., Carter, S., Sculley, D., Viégas, F. B., & Wattenberg, M., Direct–Manipulation Visualization of Deep Networks. Direct–manipulation visualization of deep networks. arXiv preprint arXiv:1708.03788, 2017.
24. Minsuk Kahng, Nikhil Thorat, Polo Chau, Fernanda Viégas, and Martin Wattenberg. GAN Lab: Understanding Complex Deep Generative Models using Interactive Visual Experimentation. IEEE Transactions on Visualization and Computer Graphics, 25(1) (VAST 2018), Jan. 2019.
25. Friedman, Jerome H. Greedy function approximation: a gradient boosting machine. Annals of statistics, 2001.
26. Goldstein, A., Kapelner, A., Bleich, J., & Pitkin, E., Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation. Journal of Computational and Graphical Statistics 24.1, 2015.
27. Cheng, Furui, Yao Ming, and Huamin Qu. DECE: Decision Explorer with Counterfactual Explanations for Machine Learning Models. IEEE Transactions on Visualization and Computer Graphics 27.2, 2020.
28. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P., SMOTE: synthetic minority over–sampling technique. Journal of artificial intelligence research 16, 2002.
29. Kliegr, Tomáš, Štěpán Bahník, and Johannes Fürnkranz. A review of possible effects of cognitive biases on interpretation of rule–based machine learning models. Artificial Intelligence, 2021.

30. Shneiderman, Ben. The eyes have it: A task by data type taxonomy for information visualizations. *The craft of information visualization*. Morgan Kaufmann, 2003.
31. Amar, Robert, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. *IEEE Symposium on Information Visualization*, 2005.
32. Aamodt, Agnar, and Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI communications* 7.1 1994.
33. Shi, Sheng, Xinfeng Zhang, and Wei Fan. A modified perturbed sampling method for local interpretable model-agnostic explanation. *arXiv:2002.07434*, 2020.
34. Kaggle, Insurance Premium Prediction, Kaggle(<https://www.kaggle.com/noordeen/insurance-premium-prediction>), 2019.
35. Ahn, Jae-wook, Catherine Plaisant, and Ben Shneiderman. A task taxonomy for network evolution analysis. *IEEE transactions on visualization and computer graphics* 20.3 2013.
36. Kaggle, Pima Indian Diabetes, Kaggle(<https://www.kaggle.com/uciml/pima-indians-diabetes-database/metadata>), 2016.

37. Steed, C. A., Shipman, G., Thornton, P., Ricciuto, D., Erickson, D., & Branstetter, M., Practical application of parallel coordinates for climate model analysis. *Procedia Computer Science* 9, 2012.
38. Li, X., et al. A debiased MDI feature importance measure for random forests. arXiv 2019.



Abstract

Interpretable machine learning supports people to understand the behavior of machine learning systems. When the relationship between input data and model decisions can be well explained, advantages such as scientific understanding, stability identification, and confidence can be obtained. As the fields to which machine learning is applied to become more diverse and the demand for model interpretation increases, the subject of machine learning model interpretation is expanding from machine learning experts to non-experts. Accordingly, the need to develop a tool capable of effectively analyzing data for model analysis has increased. Visual analytics using interactive graphics assists users including both experts and non-experts easily analyzing the dataset. Therefore, the introduction of visual analytics has great potential in machine learning interpretation. This study suggests a visual analytics system that interprets the relationship between input data and output results from the machine learning model to users easily interpret the machine learning model. This study introduced two techniques to overcome the limitations of existing machine learning analysis research. First, the input data value of the dataset was changed within a reasonable range, and the trend of change in the model prediction result was tracked accordingly. Second, to increase the usability of the analysis system, data visualization techniques such as parallel coordinates and scatter plots and interactions such as filtering, grouping, and sorting were introduced into the user interface. The visual analytics system proposed by this study takes an approach to effectively interpret machine learning models through repetitive interactive

procedures that can filter machine learning results according to input variables, target variables, and predictive values. Using this system, users find insight into the complex behavior of the machine learning model and scientific understanding of input variables, target variables, and model predictions. It also assists users to understand the stability and reliability of the machine learning model. Use-case analysis explained whether the proposed system could help achieve the goal for model analysis. Furthermore, through in-depth user interviews, the effect of providing proposed visualization and interaction techniques on system usability and model interpretation ease was evaluated. Through use case analysis and in-depth user interviews, it was confirmed that tasks occurring in the interpretable machine learning field could be performed more easily and quickly, and high-level insights could be derived by the visual analytics system suggested by this research.