https://doi.org/10.7840/kics.2023.48.11.1509

비 전문가를 위한 시각화 솔루션에서의 분류 모델링 서비스 구현

엄 태 창◆, 박 민 호°

Implementation of Classification Modeling Services in Visualization Solutions for Non-Professionals

TaeChang Eom◆, Minho Park°

요 약

정보기술 인프라와 분석 기술의 지속적인 발전으로 데이터 분석이 현실 업무에 적용되어 업무 효율성을 높이고 있다. 요즘과 같이 변동성이 큰 환경에서 즉시적인 데이터 분석을 통한 현실 업무에의 적용이 조직이나 기업의 성과를 개선하는데 큰 요인이 된다. 그러나, 데이터 분석 기술이라는 높은 진입 장벽으로 인해 업무 이해도가 높은 비 전문가의 직접적인 데이터 분석 수행은 어려운 현실이다. 이에 시각화 솔루션 기반으로 데이터 분석 기술을 융합한 서비스 구현으로 비 전문가도 분류 모델링을 수행할 수 있는 기반을 제공하는데 있다.

키워드 : 시민 데이터 과학자, 기계학습, 분류, 모델링 서비스, 시각화
Key Words : Citizen Data Scientist, Machine Learning, Classification, Modeling service, Visualization

ABSTRACT

With the continuous development of information technology infrastructure and analysis technology, data analysis is applied to real-world tasks to increase work efficiency. In today's volatile environment, the application of immediate data analysis to real work is a major factor in improving the performance of organizations or companies. However, due to the high entry barriers of data analysis technology, it is difficult for non-experts with high work understanding to perform direct data analysis. Accordingly, it is to provide a foundation for non-experts to perform classification modeling by implementing a service that combines data analysis technology based on a visualization solution.

Ⅰ. 서 론

1.1 연구 배경

많은 조직과 기업에서 데이터 분석을 통해 업무 의사결정에 도움을 받고 있으며, 빠른 환경 변화의 대응이 필요한 데이터 분석을 필요로 하고 있다.

일반적으로 데이터 과학자와 업무 담당자의 협업을 통해 데이터 분석 업무가 수행 된다. 하지만, 협업으로 인한 분석 목적 이해 및 분석 업무 수행에 많은 소요시간과 고비용 문제가 발생한다.

2015년 Gartner에서는 "수학이나 통계에 대한 깊은 지식 없이도 자신의 전문지식과 데이터 과학의 원리를

※ 이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2023R1A2C1005461).
◆ First Author : Soongsil University Department of IT Convergence, francis.eom@gmail.com, 학생회원
° Corresponding Author : Soongsil University Department of IT Convergence, mhp@ssu.ac.kr, 종신회원
논문번호 : 202306-119-C-RN, Received June 8, 2023; Revised August 18, 2023; Accepted August 31, 2023

결합할 수 있는 비즈니스 사용자"를 시민 데이터 과학자 (Citizen Data Scientist)라고 정의했다.

관련 선행 연구[1,2]에서는

"데이터 과학자는 분석 이론 및 분석 기술은 있으나, 분석 데이터를 명확히 이해하고 분석을 수행하는 데의 한계"

"분석 데이터에 대한 이해력만 갖춘 비전문가도 쉽게 분석이 가능한 시스템의 요구"

"시민 데이터 과학자는 사용하기 쉬운 분석 도구 및 기술의 도움으로 예측과 같은 데이터 분석과 비즈니스 모델을 만드는 역할"로 시민 데이터 과학자의 필요성과 역할을 수행하기 위한 방법론이 제시 되었지만, 실제 서비스 구현은 후속 연구 과제로 정의 하였다.

이에 비 전문가도 쉽게 사용할 수 있는 시각화 솔루션 기반의 분류 모델링 서비스 구현 방안을 제시한다.

1.2 현재 업무 환경의 문제점

조직의 분석 업무는 그림 1과 같이 업무 담당자와 데이터 분석가의 협업으로 진행된다.

업무 담당자가 분석의 개요 및 목적을 데이터 분석가에 공유하고, 데이터 분석가는 이를 바탕으로 데이터 수집, 처리 및 모델 학습 작업을 진행한다.

이때, 데이터 분석가의 해당 업무의 지식과 분석 데이터의 이해도에 따라 많은 작업의 소요시간과 데이터 분석가에 대한 고비용 문제가 발생한다. 이로 인해 변경 대응에 대한 적절한 기회를 상실 할 수 있다.

조직 내에서는 현업 담당자의 데이터 분석 업무 교육을 통해 이를 보완하고자 하나, 분석 기술의 진입 장벽에 의해 수월하지 않은 것이 현실이다.

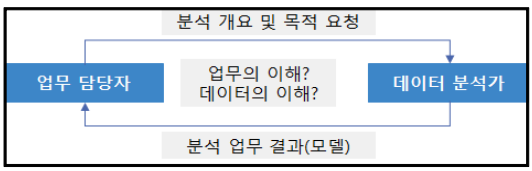


그림 1. 현 업무 환경의 문제점
Fig. 1. Problems with your current work environment

1.3 대안 서비스 제시

최근의 현업 담당자는 시각화 도구를 활용하여, 업무 데이터의 빠른 파악 및 탐색 기술은 확보 되었다.

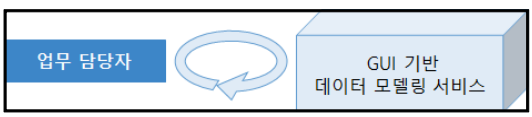


그림 2. 대안 서비스
Fig. 2. Alternative Services

이에 코딩이 필요 없는 GUI 기반의 모델링 서비스를 제시 및 구현하여, 그림 2와 같이 제안된 서비스를 활용하여 업무 담당자로 하여금 독자적 정형 데이터 기반의 다중 클래스 분류 모델링 업무가 가능 하도록 연구 한다.

Ⅱ. 관련 연구

2.1 관련 연구 방향

GUI 기반 데이터 모델링 서비스 구현을 위해 특정되지 않은 데이터의 사용과 모델링 성능 확보가 가능한 모델 학습 방안을 연구한다.

2.1.1 시각화 솔루션의 활용

시각화는 데이터를 시각적으로 표현하여 인사이트 도출 및 효과적인 전달을 가능하게 한다. 시각화는 데이터 분석 영역에서도 분석 데이터의 특성 및 분석 결과를 표현 하는데 많이 사용된다.

제안된 서비스 구현을 위해 사용되는 TIBCO Spotfire[3]는 상용 시각화 데이터 분석 솔루션으로 국내외 많은 기업에서 사용되고 있다. TIBCO Spotfire Analyst라는 저작 도구로 시각화 분석 라이브러리를 생성 및 공유 가능하여 업무 효율성을 제고 하고 있다.

시각화 분석을 위한 다양한 데이터 소스의 사용과 데이터 전처리 기능이 제공되고, 내장된 다양한 시각화 차트와 시각화 차트 간의 연계 분석을 지원한다. 최근 버전부터 Python 인터프리터가 내장되어 솔루션 내에서도 Python 라이브러리를 이용한 데이터 분석이 가능한 환경이 구성되었다.

2.1.2 분류 학습

머신러닝의 분류(Classification)는 지도학습의 한 종류로 특징 데이터를 통해 분류 Class 예측을 목적으로 한다. 현실 업무에서는 업무 처리 방안(Class)을 정의하고, 처리할 데이터를 통해 최적화 된 방안(Class)으로 분류하고 처리한다. 가령, 위험등급을 상, 중, 하 Class로 정의 하고, 요청 된 특징 데이터로 위험등급을 예측하고 결과를 업무에 활용한다. 분류 학습은 종속변수의 이진 또는 다중 범주에 따라 학습 방법이 다르며, 일반적으로 많이 사용되는 분류 학습 알고리즘으로는 Random Forest[4], KNN[5] 등 이 있다.

2.1.3 비대칭 데이터 문제

모델 학습 과정에서 종속변수의 클래스 비율에 의해 모델 예측 성능을 저해 할 수 있다. 이러한 문제를 비대칭 데이터 문제[6] (imbalanced data problem)라고 한다.

모델의 정확도는 높아도, 낮은 비율의 클래스는 재현율이 상대적으로 많이 낮아진다. 이 문제는 학습 데이터의 소실이 없는 Over Sampling 기법이나 클래스의 가중치를 적용하는 Cost-sensitive Learning 기법으로 보완 가능하다.

2.1.4 학습 알고리즘 최적화

모델 성능 확보를 위해 Hyper 파라미터 최적화를 고려해야 한다. 학습 알고리즘마다 Hyper 파라미터를 가지고 있고, 최적 파라미터의 적용만으로도 성능 향상을 가능하게 할 수 있다.

최적화 기법으로는 Grid search, Random search, Bayesian Optimization이 있다. Grid search는 수행 시간이 오래 걸리는 단점이 있어, 학습 데이터가 작은 경우에 사용되고, Random search는 Grid search에 비해 정확도가 떨어지나 수행 시간은 상대적으로 작다. Bayesian Optimization은 이전 두 기법보다 더 나은 결과를 제공하나, Bayesian Optimization을 적용하기 위한 또 다른 자체 Hyper 파라미터가 있으며, 이를 최적화해야 한다.[7]

본 서비스에서는 Random search 기반으로 범위를 좁혀가며 수 회 수행하여 최적 파라미터를 찾아 적용하는 기법을 활용한다.

2.1.5 특정되지 않은 학습 데이터

Scikit-Learn 패키지의 make_classification 함수로 다양한 가상 데이터를 생성하여 구현될 서비스의 실험에 사용한다. 이 함수는 종속변수의 클래스, 종속변수와 상관성이 있는 독립변수, 독립변수간의 선형 조합 형성이 가능하여 다양한 유형의 가상 학습 데이터를 활용하여 구현 서비스의 완성도를 높일 수 있다.

Ⅲ. 제안 서비스

3.1 서비스 환경

본 서비스는 시각화 솔루션 기반으로 제공 되므로, 저작 도구인 TIBCO Spotfire Analyst만으로 구성 된다. 아래 표 1은 서비스 환경 정보이다.

표 1. 서비스 환경 정보
Table 1. About the Service Environment

Case	Spec.
OS	Window 10 higher
CPU	2.0Ghz, 64bit 4core or more
RAM	16GB or more
HDD	More than 100GB of free space

3.2 서비스 흐름

일반적인 데이터 모델링 학습 절차와 유사 하며, 데이터 수집, 데이터 탐색, 데이터 전처리, 모델 학습, 모델 예측 진행 되며, 각 흐름별 피드백에 의해 이전 단계로의 회귀 및 재 진행 한다. 서비스 흐름은 그림 3과 같다.

그림 3. 서비스 흐름 개요
Fig. 3. Service Flow Overview

3.2.1 데이터 수집

솔루션에 내장된 데이터 조회 및 병합 기능이 사용된다. 사용 가능한 데이터는 데이터베이스 데이터 및 파일 데이터가 있다. 분석에 필요한 가공된 새로운 특징 변수 생성도 가능하다.

데이터베이스 데이터를 사용할 경우, 원천 데이터의 추가와 같은 변경이 발생 할 수 있다. 이를 방지하기 위해 데이터 수집을 위한 참조 범위를 명확히 해야 한다.

3.2.2 데이터 탐색 및 전처리

일반적인 분석에서는 분석 데이터 탐색 및 시각화 탐색을 위해 반드시 코딩 작업을 진행해야만 한다.

항목 선택 및 명령 단추 실행 등 사용자의 조작에 의해 시각화 기반의 탐색 정보를 동적으로 반영 가능한 API가 제공된다. 이는 사용자가 데이터 확인 및 시각화 표현을 위해 코딩과 같은 부가적인 작업이나 지식이 필요 없음을 의미한다. “데이터 탐색 템플릿 모듈” 구현으로 시각화된 탐색을 가능하게 한다.

데이터 탐색을 통해 확인 된 이상치 및 결측치는 내장 데이터 변환 기능으로 데이터 전처리를 지원 한다.

3.2.3 모델 학습

데이터 탐색과 같이 사용자는 기본적인 조작으로 모델 학습이 가능한 사용 가능한 “모델 학습 템플릿 모듈”을 구현한다. 분류 모델링 알고리즘은 KNN과 Random Forest를 적용한다.

사용자 편의성과 함께 모델 성능 확보를 위한 Random search 방식의 Hyper 파라미터 최적화 방안을 활용한다. 학습 성능 평가는 알고리즘에서 제공되는 Classification report, Confusion matrix를 시각화하여

제공한다.

학습 된 모델을 저장하여, 모델 예측에서 사용할 수 있도록 한다.

3.2.4 모델 예측

모델 학습을 통해 저장된 모델로 예측을 실행하고, KNN과 Random Forest의 예측 결과를 동시에 제공하여 결과를 비교 분석할 수 있는 시각화를 제공하는 "모델 예측 템플릿 모듈"을 구현한다.

3.3 서비스 구현

정의된 서비스 흐름 개발 구현 요소로 "데이터 탐색 템플릿 모듈", "모델 학습 템플릿 모듈", "모델 예측 템플릿 모듈"을 제시 하였고, 솔루션 API로 시각화 연동 개발 구현한다.

3.3.1 데이터 탐색 템플릿 모듈

사용자 조작에 의한 분석 데이터와 시각화의 동적 연계를 담당하는 모듈로 Iron Python 기반의 솔루션 API를 통해 구현 된다.

기능 구현 개발을 위해 사용자 정의 함수 ipy_GetLoadTableInfo, ipy_SelTable를 생성 하였고, 서비스 모듈에서 전역변수 역할을 지원하는 사용자 정의 요소인 Document Property 연동으로 구성된다.

ipy_GetLoadTableInfo은 생성 로드된 분석 데이터 셋을 인지하여, 분석 대상 데이터로 정의하여, 사용자의 분석 데이터 선택을 가능하게 한다.

ipy_SelTable은 선택된 분석 데이터를 시각화 구성 요소에 설정하여, 선택된 데이터를 기반으로 시각화에 표현된다.

3.3.2 모델 학습 템플릿 모듈

데이터 탐색 후, 모델 학습을 위한 모듈로 모델 학습용 사용자 정의 Python 함수를 구현하고, 솔루션 내장 Python 인터프리터를 통해 실행된다.

모델 학습 개발을 위해 참조한 Python 라이브러리는 표 2와 같다.

기능 구현 개발을 위해 사용자 정의 함수 KNN_Model[8], RF_Model[9]를 생성 하였다. 각각 KNN 분류 알고리즘과 Random Forest 분류 알고리즘을 적용한 모델 학습 함수 이다. 함수의 수행 흐름은 학습 데이터의 범주형 변수를 인코딩하고, 학습 비대상 컬럼을 제거 후 8:2 비율로 Train 데이터와 Test 데이터를 분할한다. Hyper 파라미터 최적화를 위한 Random search를 수행 후, 최적 파라미터로 최종 학습을 수행하고, 결과를 반환한다.

표 2. 참조 Python 라이브러리
Table 2. Reference Python Library

Reference Python Library	Note
numpy	Data Processing Reference
pandas	Data Processing Reference
datetime	Time Reference
sklearn.preprocessing.LabelEncoder	Categorical Data Encoding Reference
sklearn.preprocessing.OneHotEncoder	Categorical Data Encoding Reference
sklearn.model_selection.train_test_split	Data Segmentation Reference
sklearn.model_selection.RandomizedSearchCV	Random Search Reference
sklearn.neighbors.KNeighborsClassifier	KNN Reference
sklearn.ensemble.RandomForestClassifier	Random Forest Reference
sklearn.metrics.confusion_matrix	Model Results Reference
sklearn.metrics.classification_report	Model Results Reference
pickle	Model Results Reference

모델 성능 향상을 위해 Random search[10] 기법을 적용하였다. Random Forest의 파라미터 중 성능 향상에 영향이 많은 결정 트리의 개수 (n_estimators) 인자 값의 범위를 정의하고, 5회 수행하여 최적의 결과를 도출하는 인자를 최종 적용하였다. 또한, 기존 학습된 모델에서의 적용 값을 기반으로 재적용 하여 가능한 최적의 결과를 가지도록 하였다.

모델 학습 결과 지표인 Train score, Test score, Confusion matrix, Classification report를 반환하여 시각화 표현 연계 한다.

모델 학습이 완료되면 Python pickle 라이브러리로 모델을 저장하고, 인코딩 정보 등 모델 저장 정보를 반환하여 예측에서 활용 한다.

3.3.3 모델 예측 템플릿 모듈

모델 학습을 통해 저장된 모델 정보와 예측 데이터로 예측을 수행하고, 예측 결과인 predict와 predict_proba 정보를 반환한다. 기능 구현 개발을 위해 사용자 정의 함수 Predict를 생성 하였다.

3.4 서비스 사용자 인터페이스 구성

서비스 흐름에 의해 구성된 사용자 인터페이스는 7개로 사용 가이드, 데이터 명세, 데이터탐색 - 단일 변수, 데이터탐색 - 다중 변수, 모델학습-KNN, 모델학습 - Random forest, 모델 예측으로 제공된다.

3.4.1 사용가이드

사용자 인터페이스: 사용가이드는 서비스 사용을 위한 가이드 정보 제공 화면으로 분석 흐름과 각 사용자 인터페이스별 진행절차 및 주요 확인 사항에 대한 정보를 제공한다.

3.4.2 데이터 명세

서비스의 시작 지점으로 분석 하고자 하는 데이터의 기본 정보를 확인 한다. 그림 4는 사용자 인터페이스: 데이터 명세로 분석 데이터의 선택과 종속변수의 선택

그림 4. 사용자 인터페이스: 데이터 명세
Fig. 4. User interface: Data Specification

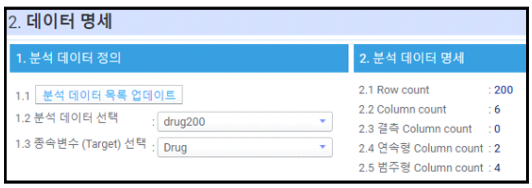


그림 5. 데이터 명세 확인
Fig. 5. Check data specification

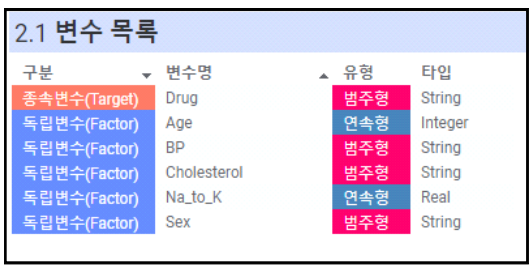


그림 6. 변수 목록
Fig. 6. List of variables

그림 7. 변수 목록
Fig. 7. Basic statistics for variables

이 주된 조작이다.

데이터 명세의 상세 내역으로 그림 5는 분석 데이터와 종속변수의 선택이다. 선택으로 분석 데이터의 명세를 확인할 수 있다.

선택된 데이터의 변수 목록과 변수 타입, 유형, 변수 구분이 그림 6과 같이 표현되며, 그림 7은 각 변수 유형에 의한 기초 통계량 정보를 제공한다.

3.4.3 데이터 탐색-단일변수

그림 8 사용자 인터페이스: 데이터 탐색-단일변수 화면으로 분석 데이터의 단일 변수 기준으로 데이터를 탐색한다. 단일 연속형 변수 또는 단일 범주형 변수 선택 조작으로 선택된 변수의 정보 확인이 가능하다.

변수 유형에 따라 표현되는 시각화는 구분되며, 시각화 간의 연계 분석이 가능하다. 그림 9와 같이 표현하고자 하는 연속형 변수와 범주형 변수를 선택 할 수 있다.

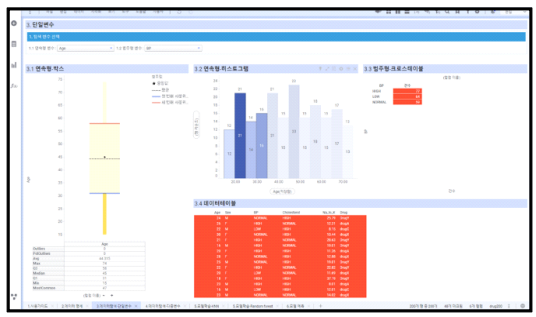


그림 8. 사용자 인터페이스: 데이터 탐색-단일변수
Fig. 8. User interface: Data Discovery - Single Variable

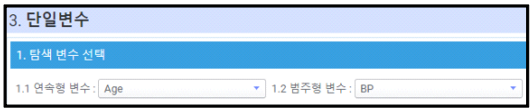


그림 9. 탐색 변수 선택
Fig. 9. Select variables

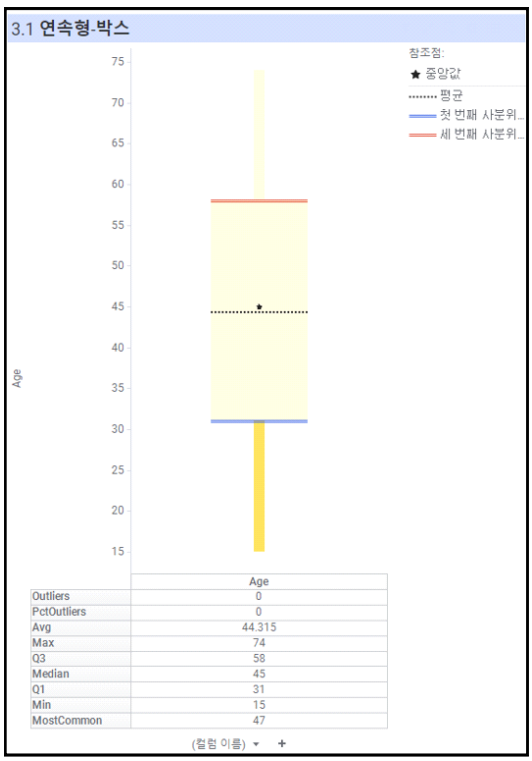


그림 10. 연속형 변수
Fig. 10. Continuous variable

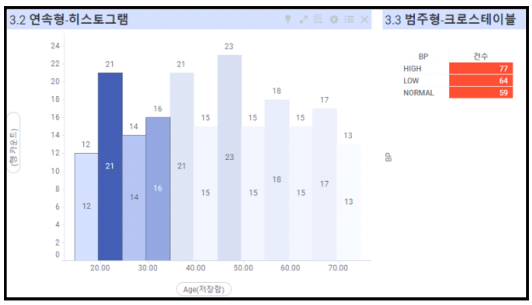


그림11. 연속형 변수와 범주형 변수
Fig. 11. Continuous / categorical variable

연속형 변수는 박스 그래프와 히스토그램을 이용하여 4 분위수 기반의 분포 및 이상치를 탐색한다. 범주형 변수는 크로스테이블로 범주별 정보를 탐색한다. 시각화 상에서 특정 지표나 범위를 선택하면 테이블 형태의 선택된 지표의 전체 데이터 연계 탐색이 가능하며, 그림 10과 그림 11과 같이 표현된다.

3.4.4 데이터 탐색-다중변수

다중 변수의 탐색을 위해 탐색 시각화의 X축, Y축을 사용자가 정의 할 수 있으며, 이를 통해 다중 변수간의 특성을 시각화된 분석을 한다. 그림 12 사용자 인터페이스: 데이터 탐색-다중변수 화면이다.

그림 13과 같이 시각화에 설정 가능한 X축, Y축 선택 항목이 제공되며, 선택 조작만으로 탐색이 가능하다.

박스 그래프는 X축을 범주형으로, Y축을 연속형으로 설정하여 범주별 이상치 및 분포를 탐색한다. 산점도는 X축, Y축을 모두 연속형으로 설정하여, 두 변수간의 분포를 탐색한다. 크로스 테이블과 히트맵은 X축, Y축을 모두 범주형으로 설정하여 두 변수간의 요약 정보

그림 12. 사용자 인터페이스: 데이터 탐색-다중변수
Fig. 12. User interface: Data Discovery - Multiple Variables

그림 13. 다중 변수 선택
Fig. 13. Select multiple Variables

그림 14. 산점도, 크로스테이블, 히트맵 표현
Fig. 14. Scatterplot, crosstable, heatmap representation

및 분포 정보를 탐색 한다. 그림 14는 산점도, 크로스테이블, 히트맵 표현이다.

3.4.5 모델학습-KNN / 모델학습-Random forest

데이터 탐색 이후, 모델 학습 실행하는 화면으로 모델명, 학습 대상 제외 변수와 선행 참조 모델을 설정으로 사용된다.

그림 15 사용자 인터페이스: 모델학습-KNN은 모델 학습 화면이고, 그림 16과 같이 학습 모델의 기본 설정으로 모델명과 학습제외 변수를 정의 할 수 있다. 학습 대상 제외 변수 옵션은 분석 데이터에는 존재하지만, 학습에는 필요 없다고 판단되는 경우 설정한다. 참조 학습 설정은 이전 학습 모델의 결과를 기반으로 Random Search의 범위를 참조 설정 할 수 있다.

실행 결과는 훈련 스코어, 테스트 스코어, 수행 parameter 정보를 표현하며, Confusion matrix와 그림 17과 같이Classification report를 시각화 표현하여 학습 실행 결과를 확인 할 수 있다.

그림 18 사용자 인터페이스: 모델학습-Random forest는 학습 알고리즘 실행 화면으로 KNN 학습과 일

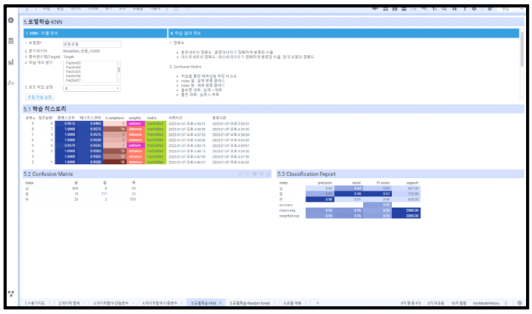


그림 15. 사용자 인터페이스: 모델학습-KNN
Fig. 15. User interface: Model Learning - KNN

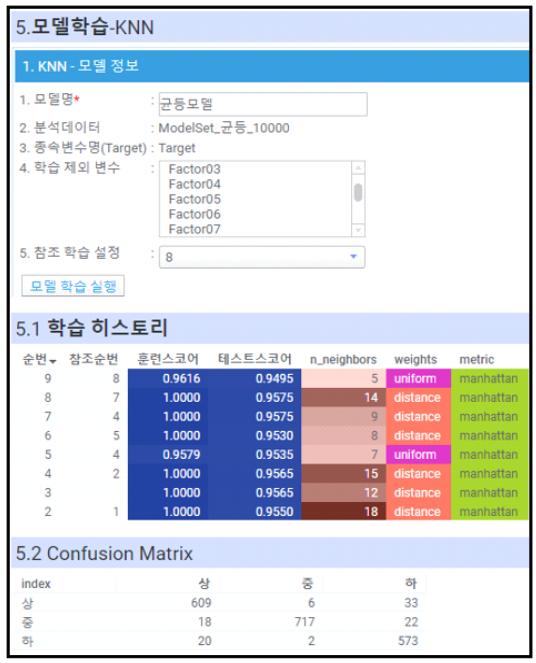


그림 16. KNN 모델 학습 설정과 학습이력 및 Confusion matrix
Fig. 16. KNN model learning settings, Learning history and Confusion Matrix

그림 17. KNN classification report
Fig. 17. KNN classification report

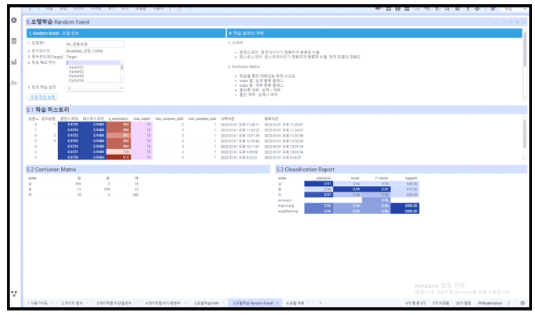


그림 18. 사용자 인터페이스: 모델학습- Random forest
Fig. 18. User interface: Model Learning - Random forest

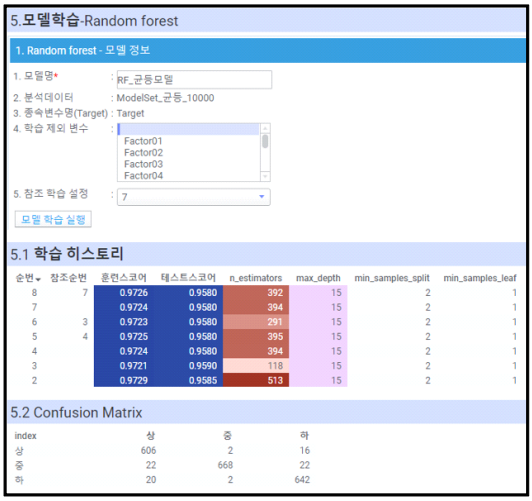


그림 19. Random forest 모델 학습 설정과 학습이력 및 Confusion matrix
Fig. 19. Random forest model learning settings, Learning history and Confusion Matrix

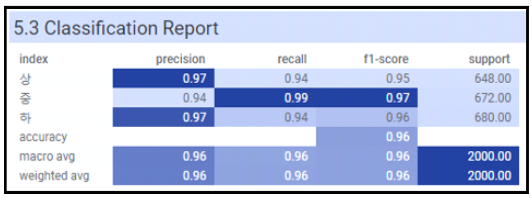


그림 20. Random forest classification report
Fig. 20. Random forest classification report

관성 있도록 구성 하였다. 그림 19, 그림 20은 KNN의 설명과 동일 하다.

3.4.6 모델 예측

그림 21은 사용자 인터페이스: 모델예측 화면으로 저장된 학습 모델과 분류 예측을 실행한다. 실행 결과는 알고리즘 별 예측 결과와 예측 확률을 제공 한다.

그림 22와 같이 모델 예측을 위한 각 알고리즘 별 학습된 모델을 선택한다.

그림 23은 선택된 모델을 통한 예측의 결과를 표현한다. 이때 KNN과 Random forest의 예측 분류 클래스로 두 알고리즘간 동일한 예측 또는 다른 예측 분포를 확인할 수 있고, 그림 24는 예측 결과 확률의 분포 이다.

그림 25는 예측 데이터와 예측 분류 결과, 예측 확률에 대한 데이터 테이블이다. 이를 통해 알고리즘간의 예측결과 차이 및 예측 확률의 높고 낮음을 비교 분석하여, 예측 결과가 동일한 클래스로 분석되면 상대적으로 정확성이 높다고 판단 할 수 있으며, 예측 확률이 낮은 수준에서 예측이 이루어 질 경우 학습 모델의 성능이 낮다는 것으로 판단할 수 있다.

그림 21. 사용자 인터페이스: 모델 예측
Fig. 21. User interface: Model prediction

그림 22. 모델 예측 설정
Fig. 22. Model prediction setting

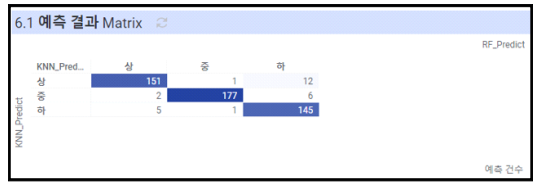


그림 23. 모델 예측 결과 Matrix
Fig. 23. Model prediction results matrix

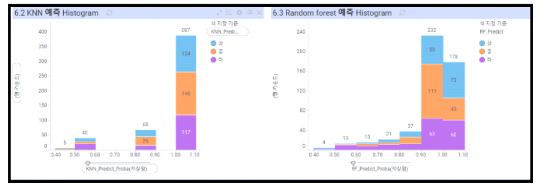
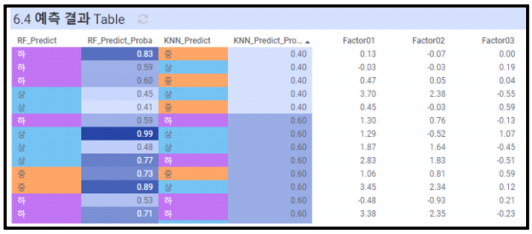


그림 24. 모델 예측 결과 히스토그램
Fig. 24. Model prediction results histogram


6.4 예측 결과 Table

RF_Predict	RF_Predict_Proba	KNN_Predict	KNN_Predict_Pro...	Factor01	Factor02	Factor03
하	0.83	중	0.40	0.13	-0.07	0.00
하	0.59	상	0.40	-0.03	-0.03	0.19
하	0.60	중	0.40	0.47	0.05	0.04
상	0.45	상	0.40	3.70	2.38	-0.55
상	0.41	중	0.40	0.45	-0.03	0.59
하	0.59	하	0.60	1.30	0.76	-0.13
상	0.99	상	0.60	1.29	-0.52	1.07
상	0.48	상	0.60	1.87	1.64	-0.45
상	0.77	하	0.60	2.83	1.83	-0.51
중	0.73	중	0.60	1.06	0.81	0.59
중	0.89	상	0.60	3.45	2.34	0.12
하	0.53	하	0.60	-0.48	-0.93	0.21
하	0.71	하	0.60	3.38	2.35	-0.23

그림 25. 모델 예측 결과 Table
Fig. 25. Model prediction results table

3.5 서비스 입출력

서비스 구현 방안으로 구현된 사용자 인터페이스별 입/출력 정보를 정리하면 표 3과 같다.

표 3. 서비스 입출력 정보
Table 3. Service Input/Output Information

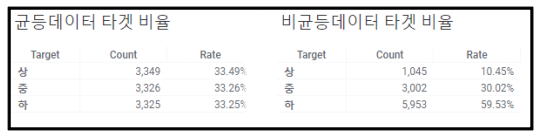
Input	Output
Data specification	
Select analysis data	1. Data specification: row count, column count, variable list, missing presence variable list 2. Basic statistics: continuous variable basic statistics, categorical variable basic statistics
Select target variable	1. Settings of the target variable
Data discovery - Single variable	
Select continuous variable	1. Boxplot 2. Histogram 3. Data table
Select categorical variable	1. Cross table 2. Data table
Data discovery - Multiple variables	
Select X-axis, Y-axis variables	1. Box plot 2. Scatterplot 3. Cross table 4. Heat Map 5. Data table
Model learning-KNN / Model learning-Random forest	
Input model name Select analysis exclusion variables Select a leading reference model Learning execution	1. Train / Test Score 2. Confusion Matrix Table 3. Classification Report Table 4. Trained Model
Model prediction	
Select predictive data Predictive execution	1. Predict / Proba. Cross table 2. Predict / Proba. Histogram 3. Predict / Proba. Table

Ⅳ. 서비스 실험 및 서비스 설문

4.1 실험 데이터 생성

제안 서비스 실험을 위해 sklearn 라이브러리의 make_classification 함수로 가상데이터를 생성 및 활용하였다. 가상데이터는 분류 학습을 위한 상, 중, 하 3개의 분류로 구성된 종속변수와 15개의 독립변수로 구성하였다. 비균등 분류 클래스 상황의 실험을 위해 분류 클래스의 비율이 다른 2종의 데이터셋으로 구성된다. 첫 번째 셋은 종속변수의 분류 비율이 균등하고, 두 번째 셋은 각각 10%, 30%, 60% 비율로 설정하였다. 그림 26은 실험 데이터로 생성된 균등 데이터와 비균등 데이터의 종속변수 비율 정보이다.

실험 데이터는 15개의 독립변수로 구성되고, 이중 4개의 독립변수는 종속변수와 상관성이 있도록 설정 하였고, 가상 데이터 중 500건을 분할하여 모델 예측 모듈 실험에 활용 하였다.


균등데이터 타겟 비율

Target	Count	Rate
상	3,349	33.49%
중	3,326	33.26%
하	3,325	33.25%

비균등데이터 타겟 비율

Target	Count	Rate
상	1,045	10.45%
중	3,002	30.02%
하	5,953	59.53%

그림 26. 실험 데이터 종속변수 비율
Fig. 26. Experimental data dependent variable ratio

4.2 실험 결과

Random forest와 KNN으로 균등/비균등 데이터로 모델 학습을 실험하였다. 균등 / 비균등 조건에서 정확도와 정밀도는 0.94이상으로 학습 되었으나, 비균등 조건에서의 재현율이 상대적으로 낮음이 확인 되었다. 실험 결과는 아래 표 4와 같이 정리할 수 있다. 정밀도, 재현율, f1-score는 분류 클래스별 최소값이다.

표 4. 실험 결과 정리
Table 4. Summary of Experimental Results

Learning algorithms	Classification Ratio	Accuracy	Precision	Recall	F1-Score
Random Forest	balanced	0.97	0.96	0.94	0.96
	imbalanced	0.97	0.97	0.83	0.91
KNN	balanced	0.95	0.92	0.92	0.94
	imbalanced	0.96	0.79	0.79	0.86

4.3 서비스 설문

TIBCO Spotfire Analyst를 사용하는 업무자 10명을 대상으로 2023년 6월 3일부터 6월 5일까지 설문이 진

행되었고, 본 연구를 통해 제안된 서비스를 총 7개 항목으로 사용자 편의성, 보완점 및 활용성 관점으로 서비스의 활용 가능성과 보완사항에 대해 파악하였다.

설문 1과 설문2는 응답자의 현재 데이터 분석 업무 현황을 파악하기 위한 설문이고, 설문 3부터 설문7은 제안된 서비스 관련 설문이다.

표 5. 제안된 서비스 활용의 설문 결과
Table 5. Survey Results of Proposed Service Utilization

Survey items	Survey response	Ratio
1. Do you have experience in data analytics with Python / R or other data modeling services?	Yes	90%
	No	10%
2. What is the level of understanding of classification learning during machine learning?	High	20%
	Usually	70%
	Low	10%
3. What is the overall level of understanding of the proposed service user interface?	High	70%
	Usually	30%
	Low	0%
4. What is the user convenience of the proposed service compared to the existing analysis environment?	High	100%
	Usually	0%
	Low	0%
5. What part of the proposed service has improved user convenience compared to the existing analysis environment? (multiple)	Configure Usage Analysis Environment	40%
	User Manipulation (Zero-Coding)	80%
	Analysis and utilization of forecast results	30%
	Fast data analytics	70%
	None	0%
	ETC	0%
6. What needs to be supplemented to perform data analysis with the proposed service? (multiple)	Description for using the service	30%
	Model learning options to improve model performance	40%
	Add Learning Algorithm	40%
	None	20%
	ETC.	0%
7. What is the data analysis usability assessment of the proposed service?	High	100%
	Usually	0%
	Low	0%

응답자의 대부분은 데이터 분석 경험과 적절한 분류 학습의 이해를 가지고 있다. 사용자 인터페이스는 70% 이해하고 있어, 전반적인 서비스 흐름은 잘 구성되었다고 볼 수 있다. 사용자 편의성은 기존 환경 대비 높은 수준으로 응답 되었다. 그 주된 요인으로는 Zero-Coding 기반의 사용자 조작, 데이터 분석가와의 협업이 필요 없는 분석의 빠른 진행으로 판단할 수 있다. 제안된 서비스의 보완사항으로는 서비스 사용을 위한 설명 강화, 모델 성능 개선을 위한 모델 학습 옵션의 보완과 학습 알고리즘 추가가 응답되었다. 제안된 서비스로의 데이터 분석 업무 활용성 평가는 높은 수준으로 응답 되었다

V. 결론 및 추후 연구

5.1 결론

제안된 서비스 구현을 위한 연구 과정에서 분석 기술이 매우 빠르게 발전되고 있음을 알 수 있었다. 분류 알고리즘의 경우 품질이 확보된 학습 데이터와 기본 옵션만으로도 양질의 학습 결과를 만들어 주고 있다.

본 연구를 통해 업무 지식과 업무 데이터의 이해도는 가지고 있으나, 데이터 분석 기술의 진입장벽에 의해 직접적인 데이터 분석을 수행하지 못하는 분석 비 전문가로 하여금 GUI 기반의 분류 모델링 서비스를 통해 일정 수준의 데이터 분석을 진행할 수 있는 기반을 제공한다.

업무 이슈에 대한 적시적 대응이 가능하다는 점과 조직 관점에서 기존의 업무 담당자와 데이터 분석가의 협업으로 인한 부수적인 업무 수행 제반 비용 절감 및 조직 구성원의 핵심 역량을 강화 할 수 있다는 기대 효과를 가지고 있다.

다만, 서비스 구현을 위해 사용된 TIBCO Spotfire로 인해, TIBCO Spotfire 사용자만이 해당 서비스의 활용 가능하다는 점이 제약사항 이다.

5.2 추후 연구

서비스 실험을 통해 확인된 것과 같이 비균등 클래스 데이터 분석 조건을 비롯한 다양한 분석 조건 관점의 연구와 이미 적용된 분류 알고리즘과 진보된 분류 알고리즘에 대한 후속 연구를 통해 적절하고 세분화된 서비스 제공이 필요하다. 또한 설문을 통해 분석 비 전문가의 원활한 서비스 활용을 위한 가이드와 같은 매뉴얼의 보완이 필요하다.

References

[1] J.-Y. Chang, “Bigdata prediction support service for citizen data scientists,” *J. IIBC*, vol. 19, no. 2, pp. 151-159 Apr. 2019. (https://doi.org/10.7236/JIIBC.2019.19.2.151)

[2] 민철희, “디지털 혁신의 시대에 ‘시민 데이터 과학자 (CDS)’로 성장하기 위해 필요한 지식과 도구들,” *대한산업공학회 춘계공동학술대회 논문집*, vol. 2022-06, pp. 1685-1699, 2022.

[3] *TIBCO Spotfire Official*, Retrieved Apr. 2, 2023, from https://www.tibco.com/ko/products/tibco-spotfire

[4] J. E. Yoo, “Random forests, an alternative data mining technique to decision tree,” *J. Edu. Evaluation*, vol. 28, no. 2, pp. 427-448, 2015.

[5] E. Choi and N. Park, “Application and development of machine learning training program based on understanding K-NN algorithm,” *J. Korean Assoc. Inf. Edu.*, vol. 25, no. 1, pp. 175-184, 2021. (http://doi.org/10.14352/jkaie.2021.25.1.175)

[6] D. Kim, J. Choi, and J. Byun, “Development of evaluation metrics that consider data imbalance between classes in facies classification,” *Geophysics and Geophysical Exploration*, vol. 23, no. 3, pp. 131-140, 2020.

[7] J. H. Kim and H. Y. OH, “The methods to improve the performance of predictive model using machine learning for the quality properties of products,” *J. KIICE*, vol. 25, no. 6, pp. 749-756, 2021. (https://doi.org/10.6109/jkiice.2021.25.6.749)

[8] *scikit-learn RandomForestClassifier Reference*, Retrieved Oct. 15, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html

[9] *scikit-learn KNeighborsClassifier Reference*, Retrieved Oct. 29, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html

[10] *scikit-learn RandomizedSearchCV Reference*, Retrieved Nov. 13, 2022, from https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.RandomizedSearchCV.html

엄 태 창 (TaeChang Eom)

2001년 2월 : 경기대학교 경영정보학과 졸업

2023년 8월 : 숭실대학교 IT융합학과 석사

<관심분야> 머신러닝, 데이터 시각화

박 민 호 (Minho Park)

2000년 2월 : 고려대학교 전자공학과 졸업

2002년 2월 : 고려대학교 전자공학과 석사

2010년 2월 : 서울대학교 컴퓨터공학과 박사

2013년 3월 : 숭실대학교 교수

<관심분야> 유/무선 네트워크 관리 및 보안, 클라우드 컴퓨팅 및 네트워크 가상화, 머신러닝 기반 시스템 및 네트워크 보안