박 사 학 위 논 문

머신러닝 기법을 이용한 공기업 재무건전성 예측모델 실증연구

2020년

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

윤 혜 란

HANSUNG
UNIVERSITY

박 사 학 위 논 문
지도교수 나도성

머신러닝 기법을 이용한 공기업 재무건전성 예측모델 실증연구

A Empirical Study on the Financial Stability Prediction Model of South Korea's Public Enterprises with Machine Learning Techniques

2019년 12월 일

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

윤 혜 란

박 사 학 위 논 문
지도교수 나도성

머신러닝 기법을 이용한 공기업 재무건전성 예측모델 실증연구

A Empirical Study on the Financial Stability Prediction Model of South Korea's Public Enterprises with Machine Learning Techniques

위 논문을 컨설팅학 박사학위 논문으로 제출함

2019년 12월 일

한 성 대 학 교 대 학 원

스마트융합컨설팅학과

스마트융합컨설팅전공

윤 혜 란

윤혜란의 컨설팅학 박사학위 논문을 인준함

2019년 12월

심사위원장 _____(인)

심 사 위 원 _____(인)

심 사 위 원 _____(인)

심 사 위 원 _____(인)

심 사 위 원 _____(인)

국 문 초 록

머신러닝 기법을 이용한
공기업 재무건전성 예측모델 실증연구

한성대학교 일반대학원
스마트융합컨설팅학과
스마트융합컨설팅전공
윤 혜 란

본 연구는 4가지 머신러닝 기법(Random Forest, XGBoost, LightGBM, DNN)을 이용한 공기업 재무건전성 예측모델 실증연구를 통해 컨설팅 현장에서 머신러닝 기법을 적용할 수 있는 방안들을 모색해보고자 하였다.

불과 2-3년 사이 데이터의 디지털화를 통한 빅데이터 시장의 급성장과 컴퓨터 성능의 발달로 인해 머신러닝 기술의 발전과 함께 머신러닝 활용에 대한 대중적 관심이 급격하게 증가했다(한국IR협의회, 2019).

전통적 통계분석 기법이 표본에서 전체 집단의 의미를 탐색하는 수학적 모델을 중시한 형태라면, 머신러닝 기법은 데이터로부터 중요한 패턴과 규칙을 찾아내어 의사결정 지원 및 예측을 수행하는 모델을 구현하는 일에 초점이 맞추어져 있다. 이에 따라 인간의 직관이 아닌 데이터를 이용한 예측을 통한 의사 결정 시스템이 점점 중요해지고, 데이터를 이용한 경영 예측이 기업의 생존과 발전에 있어서 무엇보다 중요하게 여겨지고 있다(양진용, 2017).

현재 글로벌 선도기업인 구글, 아마존, MS는 모든 비즈니스 역량을 데이터 및 인공지능으로 전환하고 있다. 우리나라도 인공지능의 활용을 위해 정부 차원에서 공공·민간 분야별로 데이터를 수집 및 제공을 위한 빅데이터 플랫폼과 센터를 구축하고, 중소·벤처기업이 데이터를 활용한 새로운 서비스를 개발할 수 있도록 다양한 지원사업을 추진 중이다.

이처럼 다양한 분야에서 머신러닝에 대한 관심과 수요가 증가하고 있기에 컨설턴트들도 머신러닝이라는 새로운 방법론에 대한 이해 및 활용 가능성에 대해 살펴볼 필요가 있다. 그 동안 머신러닝 기법을 활용한 연구는 통계학, 의학, 공학, 자연과학 분야에서 주로 이루어져 왔으나, 최근 들어 사회과학 분야에서도 활용되기 시작했다(최필선, 민인식, 2018)

지난 해부터 올해까지 우리나라 주요 공공기간의 재무건전성이 악화되고 있다는 보도들이 계속 이어졌다. 한편 2012년 이후 공공기관 재무건전성 제고 방안에 대한 연구가 꾸준히 이어져 왔지만, 머신러닝을 이용한 공공기관 경영성과나 재무건전성 예측 연구는 찾아보지 못하였다. 이에 본 연구에서는 머신러닝 기법을 이용한 재무건전성 예측모델 설계해보기로 하였다.

본 연구에서는 공기업 재무건전성 예측모델 설계를 위해 K-IFRS 도입 시점인 2011년부터 2017년까지 7개년 연속 선정된 26개 공기업 재무 데이터를 활용하였다. 데이터를 각각 5개년 단위로 나누고 4가지 머신러닝 기법을 이용하여 2016년과 2017년 공기업 재무건전성 예측모델을 설계하여 도출된 요인들을 분석하였다.

본 연구의 실증분석 결과는 다음과 같다. 먼저 예측 모델 설계 결과, 4가지 머신러닝 기법 중 Random Forest와 XGboost를 이용한 공기업 재무건전성 예측모델을 만들 수 있었다. 모델별 예측에 있어서는 두 가지 모델 모두 거의 85% 이상의 예측 정확도를 보였고, XGBoost를 이용한 예측모델이 Random Forest 모델보다 예측 정확도가 더 높은 것으로 나타났다. 하지만 나머지 2가지 기법인 LightGBM과 DNN을 통해서는 예측모델을 만들지 못하였다. 이는 데이터의 수가 너무 적어 과한 학습으로 과적합 현상으로 실제 데이터에 대한 예측 시에는 오차가 커져 예측 정확도가 더 떨어지는 것으로 확인되었다.

연도별 예측에 있어서는 XGboost와 Random Forest를 이용한 공기업 재무건전성 예측모델 모두 2016년보다 2017년도의 예측 정확도가 더 낮게 나타났다. 이러한 결과는 2017년 정권 교체 영향으로 보인다. 2017년 5월 문재인 정부가 출범하면서 정책 등 사회 전반에 많은 변화가 있었다. 이러한 변화 요인은 기존 데이터를 통한 예측 정확도를 떨어뜨리는 요인이 될 수 있다.

머신러닝 기법은 변수 중요도를 제공하는데, 독립변수의 중요도 점수가 높을수록 목표변수 예측에 중요하다고 본다. 본 연구에서는 Random Forest와 XGboost 두 모델을 통해 총자산 경상이익률의 변화, 매출액 영업이익률, 매출액 영업이익률의 변화, 영업이익손실, 영업활동 현금흐름이 중요도가 높은 변수로 도출되었다. 주로 수익성과 관련된 지표들로 이는 공기업 재무건전성 예측에 주요변수가 됨을 확인하였다. 이러한 결과는 공기업의 특성상 공익성을 추구해야 하지만 공기업의 재무건전성을 높이기 위해서는 수익성을 간과할 수 없음을 시사한다. 또한 최근 우리나라 공기업 부채가 계속 늘어가고 있는 상황에서 공기업 경영실적 평가지표 내 재무관리 지표가 점수가 줄고, 주로 정성적 평가비중이 계속 늘어나고 있어 공기업 경영실적 평가지표에 대한 냉정한 재고가 필요한 시점이다.

본 연구 결과를 통해 컨설팅 현장에서 머신러닝 기법 적용 시 고려해야 할 부분들을 제시하면 다음과 같다. 먼저 머신러닝 기법에 필요한 데이터의 양은 문제의 복잡도과 머신러닝 기법에 따라 다르다. 실제 기업 현장에서 접하는 데이터들은 정제되지 않은 비선형 형태가 대부분이며, 이러한 데이터를 가지고 패턴과 규칙을 찾기 위해서는 훨씬 더 많은 데이터를 필요로 한다. 인터넷을 통해 막대한 양의 데이터가 확보되면서 빅데이터를 기반으로 한 머신러닝의 활용 가능성은 높아졌다. 하지만 현재 머신러닝은 원하는 수준의 모델에 필요한 데이터의 양이 표준화되어 있지 않다. 또한 컨설팅 현장에서도 기업의 규모나 기업이 원하는 문제의 형태에 따라 적용하는 방법론이 다르듯 기업이 가지고 있는 데이터의 규모나 형태에 따라 적절한 머신러닝 기법을 사용할 필요가 있다. 현재 머신러닝 기법은 계속 발전하고 있지만 전통적인 통계 분석 기법처럼 표준화된 형태의 방법론을 찾기 어렵다. 때문에 실제 머신러닝 기법의 적용에 있어서는 다양한 시행착오를 통한 노하우를 쌓아가야 할 것이다.

본 연구에서 공기업 재무건전성 예측모델 설계 시 정권의 변화 등 외부 요인에 대해 고려하지 못하였는데, 이는 실제 데이터 예측에 영향을 주요 변인이 될 수 있다. 따라서 머신러닝 기법을 이용한 예측모델 설계 시 외부 예측 시간에 따라 예측 대상의 특성이 변하는 것을 모니터링하여 그 변화를 반영하거나 주기적으로 재학습하는 작업이 필요하다(Žliobaitė, 2010).

한편 본 연구를 통하여 머신러닝 기법의 유용함도 확인할 수 있었다. 먼저 머신러닝 이용하여 설계한 예측모델은 동일한 형태의 새로운 데이터가 주어졌을 때 이미 설계한 코드를 활용하여 쉽게 예측 결과를 확인할 수 있다. 즉, 2016년 공기업 재무건전성 예측모델을 설계한 코드에 해당 연도 코드명만 바꾸어 입력하여 손쉽게 2017년 예측모델 결과를 확인할 수 있었다. 또한 머신러닝 기법에서 제공하는 변수 중요도를 통해 목표변수 예측에 기여하는 주요 변수들을 쉽게 확인할 수 있다. 이러한 머신러닝 기법의 장점들은 컨설팅 현장에서도 유용하게 사용할 수 있는 부분이라고 여겨진다.

이제 인공지능은 4차 산업혁명을 촉발하는 핵심동력으로 산업구조의 변화는 물론 사회제도의 변화를 불러올 것이며, 인공지능 역량이 기업의 성장 요인이 될 것이라고 전망하고 있다(한국IR협의회, 2019). 컨설턴트는 기업의 문제해결을 위한 솔루션을 제공하는 전문가로 산업 및 경영환경의 변화에 민감하게 반응하고 민첩하게 대응해야 한다. 이에 컨설턴트라면 현재 빠른 속도로 발전하고 있는 인공지능에 대한 빠른 이해와 컨설팅 현장에서의 적용 방안을 준비해야 한다.

본 연구는 전통적인 통계분석 기법이 아닌 머신러닝 기법을 이용한 예측모델 실증연구를 통해 컨설팅 현장에서 머신러닝 기법을 적용할 수 있는 방안을 모색해보았다는 점에서 의의가 있가 있다. 본 연구가 컨설턴트들에게 머신러닝의 필요와 중요성에 대한 인식과 함께 컨설팅 현장에서의 머신러닝 기법의 적용을 위한 하나의 발판이 되기를 기대한다.

【주요어】 컨설팅, 공기업, 재무건전성, 재무건전성지수, KJY Score, K-IFRS, 머신러닝, 예측분석, 예측모델, RandomForest, XGBoost, LightGBM, DNN

목 차

표 목 차

그 림 목 차

I. 서 론

1.1 연구의 배경 및 목적

1.1.1 연구의 배경

불과 2-3년 사이 인공지능, 머신러닝, 딥러닝에 대한 관심이 급격하게 증가했다. 머신러닝이 이렇게 빠른 속도로 발전하게 된 배경에는 빅데이터의 등장과 컴퓨터 성능의 향상에 있다. 인터넷에 연결된 모든 기기를 통해 막대한 양의 데이터가 확보되어, 빅데이터를 기반으로 한 인공기능의 활용 가능성이 더 높아졌다. 컴퓨터 성능의 향상, 특히 GPU(Graphics Processing Unit, 그래픽처리장치)의 발달은 이미지 인식부터 음성 인식 등의 정밀도를 높이는 데 크게 기여했으며, 딥러닝으로 훈련된 시스템의 이미지 인식 능력은 이미 인간을 앞서고 있다(한국IR협의회, 2019).

전통적 통계분석 기법이 표본에서 전체 집단의 의미를 탐색하는 수학적 모델을 중시한다면, 머신러닝 기법은 데이터로부터 중요한 패턴과 규칙을 찾아내어 의사결정 지원 및 예측을 수행하는 모델을 구현하는 일에 초점이 맞추어져 있다. 이에 따라 인간의 직관이 아닌 데이터를 이용한 예측을 통한 의사 결정 시스템이 점점 중요해지고, 데이터를 이용한 경영 예측이 기업의 생존과 발전에 있어서 무엇보다 중요하게 여겨지고 있다(양진용, 2017).

현재 글로벌 선도기업인 구글, 아마존, MS는 모든 비즈니스 역량을 데이터 및 인공지능으로 전환하고 있다. 우리나라에서도 공공기관이나 일반 기업에서 인공지능의 활용에 대한 관심과 수요가 급증하고 있다. 또한 정부의 데이터·AI경제 활성화 계획(2019)에 따르면, 공공·민간 분야별로 데이터의 수집 및 제공을 위한 빅데이터 플랫폼과 센터를 구축할 예정이며, 중소·벤처기업이 데이터를 활용하여 새로운 서비스를 개발할 수 있도록 데이터 구매 및 가공 비용을 지원하는 사업을 추진 중이다. 또한 공공데이터 개방 및 효율적으로 관리될 수 있도록 정부차원의 데이터플랫폼 구축 및 고도화를 진행하고 있다.

이처럼 다양한 분야에서 머신러닝에 대한 필요가 증가하고 있기에 컨설턴트들도 새로운 방법론에 대한 이해와 함께 머신러닝 기법의 활용 가능성에 대해 살펴볼 필요가 있다. 그 동안 머신러닝 기법을 활용한 연구는 통계학, 의학, 공학, 자연과학 분야에서는 주로 이루어져 왔으나, 최근들어 사회과학 분야에서도 활용되기 시작했다(최필선, 민인식, 2018)

1.1.2 연구의 목적

본 연구에서는 기업에서도 머신러닝에 관심이 급증하는 시점에서 4가지 머신러닝 기법(Random Forest, XGBoost, LightGBM, DNN)을 이용한 공기업 재무건전성 예측모델 실증연구를 통해 컨설팅 현장에서 머신러닝 기법을 적용할 수 있는 방안들을 모색해보고자 하였다.

지난 해부터 올해까지 우리나라 주요 공공기간의 재무건전성이 악화되고, 부채가 증가하고 있다는 보도가 계속 이어졌다. 2012년 이후 공공기관 재무건전성 제고방안에 대한 연구가 꾸준히 이어져 왔지만, 머신러닝을 이용한 공공기관 경영성과나 재무건전성 예측 연구는 찾지 못하였다. 이에 머신러닝 기법을 이용한 재무건전성 예측모델 만들어 공기업의 재무건전성을 예측할 수 있는 요인들을 찾아보고, 컨설팅 관점에서 시사점을 찾아보고자 하였다.

공기업 재무건전성 예측모델을 설계하기 위해 2011년부터 2017년까지의 7개년 공기업 재무 데이터를 사용하기로 하였다. 2011년 K-IFRS 도입으로 공공기관 및 일반기업의 회계기준 변경되었기 때문이다. 그런데 데이터셋의 규모가 작아 머신러닝 기법을 이용한 예측모델을 만들 수 있을지에 대한 우려도 있었다. 하지만 반대로 예측모델 설계 결과를 통해 도출된 요인의 특징과 예측 모델이 만들어지지 않았을 때의 원인들을 분석 정리하여 제시한다면 컨설턴트들이 실제 컨설팅하는 과정에서 머신러닝 기업을 적용하는데 더 도움이 되리라 여겨졌다.

이에 본 연구의 목적은 4가지 머신러닝 기법을 이용한 공기업 예측모델 실증연구를 통해 컨설턴트들에게 머신러닝 기법의 필요한 대한 인식 제고와 컨설팅 현장에서 머신러닝 기법을 적용하는 방안을 모색하는데 있다.

1.2 연구의 방법 및 논문의 구성

1.2.1 연구의 방법

본 연구의 목적을 달성하기 위하여 문헌을 통한 연구와 조사한 내용을 바탕으로 한 실증분석을 진행하였다.

먼저 이론적 배경과 선행 연구 내용은 국내외 다양한 논문 및 학술저널, 연구보고서 및 단행본 그리고 인터넷을 통하여 다양한 선행연구 자료들을 수집하였다. 자료는 주로 구글(Google), 구글 학술(https://scholar.google.com), 학술연구정보서비스(RISS), 학교 학술정보관 사이트, 기관 사이트의 검색을 통해 수집하였고, 도서관 및 서점을 통하여 이론적 배경에 관한 자료들을 확보하였다.

공기업 재무건전성 예측 모델 구축 및 검증을 위해 필요한 데이터인 공기업 재무자료(결산서)는 기획재정부 출자관리과에 정보공개 요청을 통해 얻었으며, 일부 자료는 기관 사이트를 통해 확보하였다.

본 연구에서는 공기업 재무건정성 예측모델 설계를 위해 목표변수를 재무건전성 1개로 설정하고, 독립변수는 공기업 재무 데이터를 활용하여 총 27개를 설정하였다. 목표변수인 재무건전성은 총 11개 지표를 가지고 산출한 재무건전성지수인 KJY Score[1)]로 설정하였다. 이어 독립변수는 공기업 결산서 내 재무제표 계정과목과 계정과목을 가지고 일부 파생변수를 만들어 총 27개를 설정하였다.

본 연구에서는 예측모델 설계를 위해 최근 많이 사용되고 있는 Random Forest, XGboost, LGBM, DNN 4가지 머신러닝 기법을 이용하였다. 모델 설계를 위해 주피터 노트북(Jupyter Notebook) 툴을 이용하고, 파이썬(Python) 프로그래밍 언어를 사용하여 코드를 작성하였다.

---

1) KJY Score: 11개 재무지표를 통해 산출된 재무건전성지수를 의미한다. 정준수(2017)가 2011회계연도 1개년 공공기관 재무자료를 가지고 11개 지표로 산출된 재무건전성지수를 도출하여 단순화된 지수로 공공기관 재무적 성과 측정 지표로서의 활용 가능성을 검증하였다. 이어 본 연구자(윤혜란)가 2011부터 2017년까지 7개년 회계연도 공기업 재무 데이터를 가지고 11개 지표로 산출된 재무건전성지수를 도출하여 공기업 재무적 성과에 관한 대표성을 재검증하였다. 한편 이 11개 지표로 산출에 대해 김상봉 교수가 처음 제안하고 자문을 제공하여 이 11개 지표로 산출된 재무건전성지수를 세 사람의 영문 성을 각각 따서 KJY Score라고 명명하였다.

예측 모델을 설계하기 전 데이터 전처리를 통해 결측치를 처리하고, 상관분석을 통해 변수 간 유의성 여부를 파악하였다. 이후 훈련 데이터와 테스트 데이터를 구분한 뒤 상기의 4가지 머신러닝 기법을 이용하여 공기업 재무건전성 예측모델을 설계하였다. 설계된 예측 모델을 통해 도출된 예측값과 실제값을 비교하고, 결과를 통해 도출된 요인들의 특징을 분석하여 정리하였다.

1.2.2 논문의 구성

본 논문은 총 다섯 장으로 구성되어 있으며, 각 장의 주요 내용은 다음과 같다.

제 1장은 서론 부분으로 본 논문의 연구배경 및 목적, 연구 방법 및 논문의 구성에 관한 내용을 제시하였다.

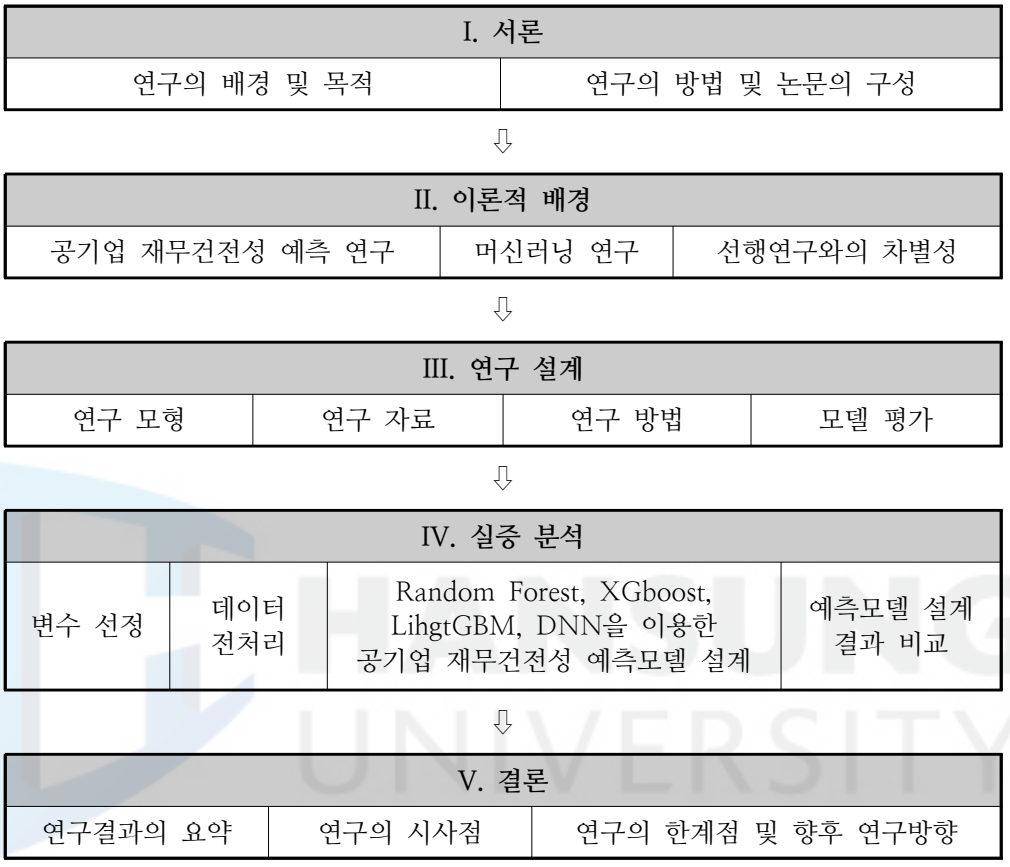
제 2장은 본 논문의 이론적 배경과 선행연구 부분으로 공기업 재무건전성 예측 연구와 머신러닝에 대한 이론적 내용, 선행연구와의 차별성에 대해 제시하였다. 먼저 공기업 재무건전성 예측 연구와 관련하여 공기업 지정 및 분류 기준을 제시하고, 공기업 재무적 성과 지표에 관한 연구와 목표변수로 설정한 재무건전성 지수에 관한 이론적 배경을 제시하였다. 이어 머신러닝의 개념과 유형에 관한 이론적 배경과 공공기관 및 기업 관련 머신러닝 기법을 이용한 선행연구 내용을 살펴보고, 본 논문에서 공기업 재무건전성 예측을 위해 사용한 4가지 머신러닝 기법(Random Forest, XGboost, LightGBM, DNN4)에 관한 내용을 제시하였다.

제 3장은 연구설계 부분으로 선행연구를 기초로 재무건전성 예측을 위한 연구 모형과 연구 자료, 연구 방법 및 모델 평가 방법에 대해 제시하였다.

제 4장은 실증분석으로 4가지 머신러닝 기법을 이용한 예측모델 설계 과정부터 결과를 제시하였다. 예측모델 설계 결과를 통해 공기업 재무건전성 예측값과 실제값 비교를 통해 예측모델 예측 정확도를 알아보았고, 결과를 통해 도출된 요인들의 특징을 분석하여 정리하였다.

제 5장에서는 본 연구결과의 요약 및 시사점을 제시하고, 연구결과의 한계점 및 향후 연구 방향을 제시하였다.

본 논문의 구성 및 흐름은 [그림 1-1]과 같다.



[그림 1-1] 논문의 구성 및 흐름도

2.1 공기업 재무건전성 예측에 관한 연구

2.1.1 공기업 지정 및 분류기준

본 연구에서의 공기업은 '공공기관의 운영에 관한 법률'에 의해 분류된 공공기관을 의미한다. 공공기관이란 정부의 투자·출자 또는 정부의 재정지원 등으로 설립·운영되는 기관으로, '공공기관의 운영에 관한 법률(이하 공운법)' 제4조 제1항 각호의 일정 요건에 해당하여 기획재정부장관이 지정한 기관을 의미한다. 다만, '공운법' 제4조 제2항에 따라 ① 구성원 상호간의 상호 부조·복리증진·권익향상 또는 영업질서 유지 등을 목적으로 설립된 기관, ② 지방자치단체가 설립하고, 그 운영에 관여하는 기관, ③ 한국방송공사와 한국교육방송공사는 공공기관으로 지정할 수 없다.

기획재정부장관은 '공운법' 제4조 제1항에 따라 다음 각 호의 어느 하나에 해당하면 공공기관으로 지정할 수 있다. 요약하면 ① 다른 법률에 따라 직접 설립되고 정부가 출연한 기관, ② 정부지원액이 총수입액의 2분의 1을 초과하는 기관, ③ 정부 단독 혹은 정부와 제1호 내지 제3호 해당 기관이 합하거나 또는 정부와 제1호 내지 제4호의 해당 기관이 합하여 100분의 50 이상의 지분 또는 100분의 30 이상의 지분을 가지고 임원 임명권한 행사 등을 통하여 당해 기관의 정책결정에 사실상 지배력을 확보하고 있는 기관, ④ 제1호 내지 제4호의 해당 기관이 합하여 설립하고, 정보 또는 설립기관이 출연을 기관이 이에 해당한다.

공공기관은 '공운법' 제5조에 따라 자체수입 비율 및 정원 기준에 따라 크게 공기업, 준정부기관, 기타공공기관 3가지로 구분되며, 공기업과 준정부기관은 다시 두 가지로 세분되고 있다[2]. 공기업은 자산규모와 자체수입액 규모를 기준으로 시장형과 준시장형으로 분류되며, 준정부기관은 기금 관리 유무를 기준으로 기금관리형과 위탁집행형으로 분류되고 있다. 기타공공기관은 공기업·준정부기관을 제외한 공공기관으로 기타공공기관으로 분류한다(국회예산정책처, 2019).

---

2) 출처: 알리오플러스. 기관정보-'공공기관이란?', http://alioplus.go.kr/organization/organByPub.do

공운법에 따른 공공기관 유형 및 분류기준 내용을 요약하면 〈표 2-1〉과 같다.

〈표 2-1〉 공공기관 유형 및 분류기준

유형		분류기준
① 공기업		직원 정원 ≧ 50인 & 자체수입/총수입 ≧ 50%
	시장형	자체수입/총수입 ≧ 85% (자산규모 2조원 이상)
	준시장형	시장형 공기업이 아닌 공기업
② 준정부기관		직원 정원 ≧ 50인 & 자체수입/총수입 〈 50%
	기금관리형	「국가재정법」에 따라 기금 관리 또는 위탁관리
	위탁집행형	기금관리형 준정부기관이 아닌 준정부기관
③ 기타공공기관		직원 정원 〈 50인 & 공기업·준정부기관이 아닌 공공기관
	연구목적 공공기관	연구개발을 목적으로 하는 기관

출처 : '공공기관의 운영에 관한 법률', 2019 대한민국 공공기관(국회예산정책처, 2019), 알리오플러스(http://alioplus.go.kr/organization/organByPub.do) 자료를 바탕으로 재작성

공운법에 따른 연도별 공기업 지정 현황은 〈표 2-2〉와 같다. 공운법 제6조에 따라 기획재정부 장관은 매 회계연도 개시 후 1개월 이내 혹은 경우에 따라 회계연도 중 공공기관을 새로 지정하거나, 지정을 해제하거나, 구분을 변경하여 지정할 수 있다.

〈표 2-2〉 연도별 공기업 지정 현황 (2011~2019년)

(단위: 개)

회계연도	2011	2012	2013	2014	2015	2016	2017	2018	2019
공기업 지정 기관 수	27	28	30	30	35	30	35	35	36

출처 : 국회예산정책처(2019). 2018회계연도 공공기관 결산분석 I

본 연구에서는 2011회계연도부터 2017회계연도까지 7개년 연속 지정된 26개 공기업 제무제표를 가지고 공기업 재무건전성 예측모델을 설계하였다.

2.1.2 공기업 재무적 성과 평가 지표 관련 연구

2.1.2.1 공공기관 경영실적 평가 지표

공운법에 따르면, 공공기관의 자율적 운영을 보장하면서 공공기관의 특수성을 고려하여 공공기관 운영상 책임성·신뢰성·투명성 등의 중요 가치가 보장되도록 여러 장치를 마련하고 있다[3].

이 중 '경영실적 평가제도'가 있다. 공운법 제48조에 따른 '경영실적평가제도'는 공기업·준정부기관의 자율·책임경영체계 확립을 위해 매년 경영노력과 성과를 공정하고 객관적으로 평가하는 제도이다. 이에 공기업·준정부기관은 법률 제47조에 따라 매년 3월 20일까지 전년도 경영실적 보고서를 제출하고, 기획재정부는 이를 기초로 하여 경영실적을 평가한다(기획재정부, 2018).

공공기관의 경영평가는 사전에 설정한 평가지표와 기준에 따라 진행되며, 평가지표와 기준은 평가대상이 되는 연도가 시작하기 전에 확정한다(정준수, 2017). 이에 따라 연도별 경영실적 평가지표가 일부 변동이 생기기도 하였다. 〈표 2-3〉은 연도별 공공기관 경영실적 평가지표를 나타낸 것이다.

〈표 2-3〉 연도별 공공기관 경영실적 평가지표 (2011~2019년)

<table>
<tr><th colspan="2">2011~2013년</th><th></th><th>2014~2017년</th><th>2018~2019년</th></tr>
<tr><th>평가범주</th><th>주요평가내용</th><th>평가범주</th><th>주요평가내용</th><th>주요평가내용</th></tr>
<tr><td rowspan="4">리더십
·
책임경영</td><td>리더십</td><td rowspan="8">경영관리</td><td rowspan="4">경영전략 및 사회공헌</td><td rowspan="3">경영전략 및 리더십</td></tr>
<tr><td>책임경영</td></tr>
<tr><td>국민평가</td></tr>
<tr><td>사회적기여</td><td>사회적 가치 구현</td></tr>
<tr><td rowspan="4">경영효율</td><td>업무효율</td><td>업무효율</td><td>업무효율</td></tr>
<tr><td>조직 및 인력자원관리</td><td>조직·인적자원 및 성과 관리</td><td rowspan="2">조직·인사·재무관리</td></tr>
<tr><td>재무예산관리 및 성과</td><td>재무예산 관리 및 성과</td></tr>
<tr><td>보수 및 성과관리</td><td>보수 및 복리후생비 관리</td><td>보수 및 복리후생비</td></tr>
<tr><td colspan="2">노사관리</td><td></td><td>혁신과 소통</td></tr>
<tr><td>주요사업</td><td colspan="4">공공기관의 주요사업별 계획·활동·성과를 종합적으로 평가
2018년부터 계량지표의 적정성 추가</td></tr>
</table>

출처: 기획재정부, 연도별(2011-2019) 공공기관 경영평가편람 내 평가지표 내용 종합 정리

3) 출처: 알리오플러스. 기관정보-'공공기관이란?', http://alioplus.go.kr/organization/organByPub.do

〈표 2-3〉에서 보는 것처럼, 2014년부터 공공기관 경영실적 평가범주가 경영관리 한 범주로 통합되었으며, 주요평가내용 항목 또한 통합 또는 변경되었다. 이후 2018년 주요평가내용 항목 일부가 다시 분리 또는 통합되었고, 주요사업 내용 중 계량지표의 적정성 내용이 추가되었다. 주요평가내용 항목의 변화와 함께 평가점수도 변동이 있었다.

이처럼 경영실적 평가지표 일부가 통합 또는 분리되거나 변경되기는 했지만, 공공기관 경영실적 평가지표의 내 '재무예산관리및성과(재무관리)'의 주요내용은 재무(예산) 안정성, 투자 및 집행효율성, 중장기 재무관리계획 이행실적, 건전한 재무구조 및 합리적 예산운용을 위한 재무예산 관리시스템 구축 및 운영 성과 평가에 대한 것이다.

종합경영평가와 관련하여 '재무예산관리 및 성과(재무관리)' 평가지표 총점은 2011년에서 2013년까지 12점이었다가 2014년 18점으로 늘었다. 이후 2015년 15점, 2016년 14점, 2017년 10점으로 점점 줄었고, 2018년과 2019년은 5점으로 대폭 줄었다. 2011년부터 2019년까지 연도별 공기업 경영실적 평가지표 재무예산관리 점수에 대한 상세 내용은 〈표 2-4〉와 같다.

〈표 2-4〉 연도별 공기업 경영실적 평가지표 재무예산관리·성과 점수 (2011~2019년)

평가지표	2011-2013년	2014년	2015년	2016년	2017년	평가지표	2018-2019년
재무예산관리	4	3	2	2	3	재무예산 운영성과	5
계량 관리 업무비	2	3	3	2	(1)		
자구노력이행 성과		6	4	4	2		
재무예산성과 (부채감축달성도) (중장기재무관리계획)	6	6 (4)	6 (2) (2)	6 (2) (2)	5 (1.5) (1.5)		(1)
합계	12	18	15	14	10	합계	5

출처: 기획재정부, 연도별(2011-2019) 공공기관 경영평가편람 내 평가지표 내용 종합 정리

〈표 2-4〉에서 보는 것처럼 연도별 공기업 재무예산관리 및 성과에 대한 평가지표 점수가 조금씩 변화가 있었다. 2011년도부터 2013년도까지 평가지표 항목과 총점이 같았으나, 2014년 자구노력이행 성과 항목이 추가되면서 총점

18점으로 늘었고, 재무예산성과 내 중장기재무관리 항목이 추가되었다. 이는 2013년 공기업의 계속된 부채 증가로 인한 정부의 정책이 반영된 결과이다. 2015년부터는 재무예산성과 내 부채감축달성도가 추가되면서 중장기재무관리 계획 평가점수가 반으로 나뉘었고, 자구노력이행성과와 계량 관리 업무비에 대한 점수가 줄면서 총점이 줄어들었다. 2018년부터는 세부적으로 나뉘던 평가지표가 하나로 통합되었으며 2017년 10점이던 점수가 절반인 5점으로 줄었다. 한편 평가지표 통합 항목 내 부채감축달성도는 구분되어 있었다.

2011년 이후 정부 부채의 지속적인 증가로 안하여 공공기관의 재무상태 및 경비 사용에 외부적인 관심이 커지고 있다. 하지만 현재 공기업의 적자 비중이 계속해서 증가하는 상황에서 공기업 경영실적 평과 내 '재무예산관리 및 성과 (재무관리)' 지표 점수 비중이 2016년 이후 절반으로 줄었다. 공기업의 특성상 공익성을 추구해야 하지만, 공기업의 재정 적자는 직접적으로 국가 부채로 연결되는 상황이기에 공기업 경영평가와 관련 평가지표별 비중에 대해 재고한 필요하다.

2.1.2.2 공기업 재무적 성과 지표 관련 선행 연구

2011년 이후 공공부채가 국가채무 규모를 넘어서면서 공공기관 재무건전성에 대한 다양한 분석과 대책에 대한 연구가 이루어지기 시작했다(박진, 최진욱, 박진희, 김지영, 허경선, 2012) 이에 본 연구에서는 공공기관 재무건전성 예측 분석에 있어 재무건전성이나 재무적 성과와 관련 어떤 지표를 사용했는지를 중심으로 선행 연구를 살펴보았다.

기업의 재무건전성을 나타내는 지표로 일반적으로 부채비율 사용한다. 부채비율은 기업 자산 중 부채가 차지하는 비율로, 기업의 재무구조를 나타내는 대표적인 지표로 기업 재무건전성 평가 목적으로 주로 사용된다(정준수, 2017) 공공기관의 결산지표와 일부 연구에서는 재무건전성 지표로 수익성, 안정성, 성장성, 생산성 및 활동성 지표 등을 제시하였다. 그 외 공공기관 재무건전성이나 재무 상태를 보는 자료는 부채비율 외 당기순이익을 포함하는 등 연구자에 따라 다양하게 나타났다.

〈표 2-5〉는 공공기관의 재무적 성과와 관련 지표 선행 연구에 관한 내용을 요약하여 제시하였다.

〈표 2-5〉 공공기관 재무적 성과 관련 지표 선행 연구(요약)

연구자 (연도)	연구내용	연구자료	연구방법	재무적 성과 관련 지표	연구결과
한창구 (2010)	“지방공기업의 효율화를 위한 경영평가지표에 관한 연구 : 시설관리공단 평가지표를 중심으로”	경기도 내 14개 공단, 2004-2006년 결산자료	평가지표별 경영평가분석	자산, 부채, 자본	경영성과 등급 비교 시 재무건전성보다 규모가 큰 기관 등급이 높게 나타나 지표체계가 기관별 고유특성 반영 못하는 것으로 추론, 가중치 조정 필요
최재훈 (2010)	“공기업 회계에 있어 K-IFRS의 효율적 도입방안에 관한 연구“	해외 사례 및 IFRS관련 문헌	문헌 연구	부채비율	IFRS 도입 시 공공기관 관리 선진화 기반 구축 기대,원가 위주에서 공정가치 평가로 재무지표 개선 효과
조택, 이창균 (2010)	“공공기관 경영평가 지표의 타당성에 관한 연구- 계량지표를 중심으로”	2009년 공공기관 계량평가 결과, 2008-2009 재무 데이터	상관분석 선형회귀분석	재무비율 (부채비율 증가율, 총자산 증가율)	공공기관 평가유형별로 계량평가 결과와 주요 재무비율 간 기준 타당성 미흡, 구성 타당성 일부 제한됨 확인
박용성, 남형우 (2011)	“공공기관의 외형적 특성이 경영평가 결과에 미치는 영향”	2007-2009 공공기관 경영평가자료	상관분석	부채비율 당기순이익	재무건전성이 경영평가에 긍정적 영향을 줌
정성호, 전항훈 (2012)	“공공기관의 재무건전성 제고 대안”	2004-2010(준) 시장형 공기업 22개 경영성과지표	지표 및 문헌 분석	부채, 당기순이익, 금융성부채, 부채 비율	부채 증가 재정건전성 악화, 경영평가 지표상 재무예산관리, 재무예산성과 자료가 전체 5%에 불과 경영평가 반영 미흡
표경호 (2012)	“에너지 관련 공공기관의 국제회계기준 도입이 재무제표에 미치는 효과”	2010년 에너지 관련 3개 공공기관 재무제표	사례 연구	부재비율, 당기순이익	IFRS적용 시 일부 공공기관 부채비율 감소, 당기순이익증가 등 재무건전성 제고됨, 공공기관 회계투명성 확보 기대
박원, 김태영 (2013)	“공기업과 준정부기관의 재무 특성에 따른 경영성과와 이익 조정에 관한 연구”	2005~2010년 공기업/ 준정부기관 재무자료	상관분석 회귀분석	재무적 특성 (부채비용, 국고보조금)	재무적 특성 중 부채비율만 경영성과(총자산이익률)에 음의 영향
안숙찬 (2014)	“공공기관의 경영평가 결과와 기관특성에 관한 연구”	2008-2012공기업/준정부기관 경영평가 자료와 재무자료	상관분석 회귀분석	재무적 특성 (수익성, 활동성, 성장성)	경영효율 범주 분석 시 재무적 특성이 유의하게 나타남. 계량/비계량 지표에서 계량지표점수가 최종 종합평가에 미치는 영향이 큰 것으로 나타남.
김주형 (2014)	“공기업 부채비율 결정요인에 관한 실증분석‘	2002~2012년 공기업	패널분석	부채비율	공기업 통한 정부정책사업추진, 공공요금 통제, 공기업의 무리한 사업 추진은 재무건전성을 약화시킴

연구자 (연도)	연구내용	연구자료	연구방법	재무적 성과 관련 지표	연구결과
조임곤 (2014)	“우리나라 공기업의 경영분석 지표 추이 변화에 대한 연구“	3개년 공기업 경영지표	경영지표 분석	안전성 (자기자본 비율, 부채비율)	시장형 공기업은 부채비율이, 준시장형 공기업은 유동비율이 경영분석지표와 상관관계보임.
임경희 (2016)	“공기업의 보수・복리후생비가 재무건전성과 경영성과에 미치는 영향”	2010-2014년 50개 공기업 재무제표	회귀분석	부채증가율, 부채비율, 차입금증가율	공기업 보수/복리후생비가 재무건전성 및 경영 성과 (자기자본순이익율) 제고 효익을 나타냄
정준수 (2017)	“재무건전성지수를 이용한 공공기관 재무건전성 제고방안”	2011년 공기업/ 준정부기관 결산서	재무건전성 지수 도출, 상관분석, 군집분석	재무건전성 지수	IFRS도입 후 공공기관 재무건전성 지수를 도출하여 단순화된 지수로서 재무적 성과 측정 지표로의 활용 가능성 검증
최현정, 표춘미 (2018)	“공기업 경영실적평가에서 재무제표의 이용 정도가 재정건전성에 미치는 영향”	2010-2014년 28개 공기업	회귀분석	총부채액, 부채비율, 부채비율증감, 매출액, 당기순이익	재무제표가 재정건전성(총부채액, 부채비율, 부채비율증감, 매출액, 당기순이익) 을 개선하는 방향으로 영향을 미치지 않음

본 연구에서는 정준수(2017)의 연구에서 제시한 재무건전성지수를 목표변수인 재무적 성과 지표로 선정하였다. 정준수(2017)는 공공기관이 K-IFRS 도입 후 공공기관 재무자료를 가지고 수정 재무건전성지수를 도출하여 단순화된 지수로서 공공기관의 재무적 성과 측정 지표로의 활용 가능성을 확인하였다. 이에 다음 절에서는 재무건전성지수 연구 내용을 제시하고자 한다.

2.1.3 재무건전성지수에 관한 연구

일반적으로 신용평가 기관에서는 기업의 부채상환 능력이나 재무건전성을 파악하기 위하여 복잡한 계량적 기법을 사용하여 신용등급을 도출한다. 때문에 일반 투자자들은 신용등급이 어떻게 산출되는지 알 수 없으며, 기업의 부채상환 능력이나 재무건전성에 대해 대부분 명확히 이해하지 못한다. 계량적 기법에 대한 이해가 부족한 일반 투자자나 기업은 신용등급을 예측하거나 검증하는데 한계가 있다. 그러므로 신용등급을 이용하는 수요자 입장에서는 기업에서 공시하는 재무제표를 이용하여 재무건전성을 쉽게 도출할 수 있는 Piotroski의 재무건전성지수의 사용이 유용하다(정준수, 2017).

재무건전성지수는 Piotroski(2000)가 간단한 지수로 기업의 재무건전성을 종합적으로 파악할 수 있도록 제시한 지표로 F-Score라고 부르기도 한다. 이 지표는 재무제표를 사용하여 수익성, 재무구조, 업무효율 세 가지 영역으로 나누어 수익성의 변화 및 부채비율 등 9가지 재무정보가 담겨 있다(Piotroski, 2000). Piotroski(2000)가 제시한 F-Score는 총 9가지 재무지표를 계산하여 각 지표별로 기준에 따라 1점 혹은 0점을 부여하여 총점 9점 만점으로 이루어져 있다.

Piotroski가 제시한 재무건전성지수 F-Score 산출식은 〈표 2-6〉과 같다.

〈표 2-6〉 Piotroski의 F-Score 산출식

산출식	F-Score = FROA +F_⊿ROA +F_CFO +F_ACCRUAL +F_⊿LEVER +F_⊿LIQUID +F_EQ_OFFER +F_⊿MARGIN +F_⊿TURN
산출 지표	F_ROA : 총자산경상이익률 F_⊿ROA : 총자산경상이익률의 변화 F_CFO : 총자산 대비 영업현금흐름 비율 F_ACCRUAL : 총자산 대비 영업현금흐름 비율과 총자산경상이익률 차이 F_⊿LEVER : 평균총자산 대비 장기부채 비율 F_⊿LIQUID : 유동비율 F_EQ_OFFER : 보통주 발행(유상증자) F_⊿MARGIN : 매출액 총이익률 F_⊿TURN : 총자산회전율

출처: 정준수(2017) 인용, 일부 수정

Piotroski의 F-Score를 구성하는 9가지 재무지표 항목들은 〈표 2-7〉과 같이 수익성, 재무구조, 업무효율성 세 가지 영역으로 구분할 수 있다.

〈표 2-7〉 Piotroski의 F-Score 구성 9가지 재무 지표

항 목	재무 지표
수익성	① 총자산경상이익률 (ROA)
	② 총자산경상이익률 변화 (⊿ROA)
	③ 총자산에 대비 영업현금흐름 비율 (CFO)
	④ 총자산 대비 영업현금흐름 비율과 총자산경상이익률 차이 (ACCRUAL = CFO - ROA)
재무구조	⑤ 평균총자산 대비 장기부채 비율(⊿LEVER)
	⑥ 유동비율 (⊿LIQUID)
	⑦ 보통주 발행(유상증자) (EQ_OFFER)
업무효율성	⑧ 매출액 총이익률 (⊿MARGIN)
	⑨ 총자산회전율 (⊿TURN)

출처: 정준수(2017) 인용, 일부 수정

Piotroski의 F-Score를 구성하는 개별 지표를 살펴보면 다음과 같다. 먼저 수익성 지표는 총 4개로 총자산경상이익률(ROA), 총자산경상이익률 변화(⊿ROA), 총자산 대비 영업현금흐름 비율(CFO), 총자산 대비 영업현금흐름비율과 총자산경상이익율 차이(ACCRUAL)다. 재무구조 지표는 3개로, 레버리지 변화(⊿LEVER), 유동비율 변화(⊿LIQUID), 유상증자 여부(EQ_OFFER)이다. 마지막으로 영업활동의 효율성을 나타내는 현금흐름 지표는 2개로 매출총이익률 변화(⊿MARGIN), 자산회전율의 변화(⊿TURN)이다(Piotroski, 2000).

F-Score를 구성하는 9개 지표는 다음과 같은 방식으로 1 또는 0이라는 점수가 부여된다. 먼저 수익성의 경우, 기업의 신용도에 관하여 양의 방향이면 이는 기업의 수익성이 개선되었거나 영업현금흐름이 증가된 것을 의미하기 때문에 1점을 부여하고 부정인 방향이면 0점을 부여한다. 영업현금흐름비율(CFO)과 경상이익율(ROA)의 차이를 ACCRUAL로 정의한다. 영업현금흐름이 경상이익보다 크면 당기 경상이익이 과소계상되어 당기의 수익성은 낮아지지만 차기 경상이익이 커질 가능성이 높다는 신호로 해석할 수 있기 때문에 1을 부여하고 영업현금흐름이 경상이익보다 작거나 같으면 0점을 부여한다. 재무

구조인 레버리지와 유동성의 경우, 레버리지 변화가 0이하이면 이를 재무구조가 개선된 신호로 보고 1점을 부여하고, 반대로 0점을 초과하면 재무구조가 악화된 신호로 보고 0점을 부여한다. 유동비율의 변화는 0점을 초과하면 단기 유동 위험성 감소 신호로 보고 1점을 부여하고, 0점 이하이면 단기 유동 위험성이 악화 신호로 보고 0점을 부여한다. 유상증자는 기업이 영업활동을 통해 조달된 내부자산으로 부채를 상환하지 못할 가능성이 높다고 보면 0점을 부여하고, 유상증자를 실시하지 않은 기업은 1점을 부여한다. 영업효율의 경우, 매출액총이익률과 총자산 회전율이 각각 전기에 비하여 높으면, 즉 변화율 0점 초과시 영업효율성이 호전된 신호로 보고 1점을 부여하고, 전기에 비하여 낮으면 0점을 부여한다. 이렇게 해서 계산된 F-Score의 범위는 최소 0점에서 최대 9점이다(Piotroski, 2000).

Piotroski는 F-Score를 이용하여 F-Score가 높은 기업들의 주식 수익률과 F-Score가 낮은 기업들의 주식 수익률을 비교하여 분석하였다. 분석 결과 Piotroski는 재무건전성지수 F-Scrore가 주가 상승률이 높거나 낮은 종목의 선택에 지표 역할을 할 수 있다는 연구 결과를 발표하였다. 특히, F-Score가 4이상 9미만일 때 F-Score 값과 주식투자에서 초과수익률 간에 양의 상관계가 있음을 증명하였다(Piotroski, 2000).

Piotroski의 F-Score는 다음과 같은 장점이 있다. 첫 번째는 다변량 분석 방법을 사용하는 경우 변수들 간 상관관계가 높으면 다중공선성의 문제들이 발생한다. 하지만 Piotroski가 제시한 F-Score는 변수 자체를 더한 종합 점수로 계산함으로써 변수들 간 상관관계를 고려하지 않더라도 판별해 낼 수 있다. 두 번째는 변수들이 단지 양의 방향이면 1점, 부의 방향이면 0점을 부여하여 변수들을 지수화할 필요가 없다. 세 번째는 0과 1 숫자는 확률과 같은 개념으로 사용할 수 있으므로 해석이 쉽다는 점이다(정준수, 2017).

Piotroski가 재무건전성지수 F-Score 모델 제시 이후 F-Score를 이용한 연구들이 이어졌다. 고덕필a(2003)은 F-Score와 회사채 신용등급 간 관계를 분석하였다. 기업의 재무건전성이 양호하면 부채상환 능력도 양호할 것이라고 예측한 후 분석 결과를 통해 F-Score와 회사채 신용등급 간 밀접한 관계가 있음을 밝혔다. 이러한 결과는 신용등급이 기업의 부채상환능력을 보여준다면

기업의 재무건전성, 즉 F-Score 값이 높은 기업일수록 기업의 신용등급도 높을 것이라고 예측할 수 있다.

신용등급은 부채상환 능력을 기호로 등급화한 것이다. 국내 신용평가회사들은 기업이 발행하는 회사채에 대한 원리금 상황 능력에 따라 AAA부터 DDD까지 단계별로 등급을 부여하고 있다. 이에 AAA부터 BBB까지를 투자등급, BB 이하는 투기등급으로 분류한다. 신용등급 분류를 통해 기업은 자기 신용능력에 따라 적정금리로 시장에서 자금을 조달할 수 있고, 금융기관은 기업의 신용등급을 이용하여 리스크 관리와 신탁 자산 운용에 필요한 정보를 제공하며, 투자자나 투자기관에게는 합리적 의사결정을 위한 객관적인 정보 제공이 가능하다. 이러한 신용등급에 관한 정보는 일반 투자자들과 기관 투자자들의 의사결정에 중요하게 작용하여 주식시장에 반영된다. 하지만, 신용평가 기관에서 제공하는 기업의 신용등급은 그 산출과정을 공개하지 않으므로 신용등급의 결정 과정에 대한 내용 이해가 어려워 기업 입장에서는 적시성 있는 신용등급 관리가 어려운 것이 현실이다(박경덕, 한길석, 윤석진, 2008) .

고덕필b(2003)은 일반 투자자는 신용등급을 쉽게 예측할 수 있고, 기업 스스로는 신용등급을 예측하고 관리하기 쉬운 방법을 제시하였다. 고덕필b(2003)은 그의 연구에서 Piotroski(2000)의 F-Score를 수정 보완한 지수를 제시하였다. 이자보상배율 지표를 추가·보완하여 10점 만정으로 구성된 재무건전성지수를 제시하였다. 재무건전성 구성 지표로 이자보상배율을 추가한 이유는 이자보상배율이 기업의 영업활동 수입으로 차입금과 회사채 이자를 정상적으로 갚을 수 있는지 여부를 알 수 있는 유용한 재무정보이기 때문이다.

김형민(2003)에 따르면 신용등급이 높은 기업이 재무건전성지수도 높으며, 신용등급과 재무건전성지수 간 뚜렷한 양의 관계가 있다. 김형민(2003)은 기업의 F-Score가 6점 이상이면 투자적격등급, 6점 이하이면 투기등급에 포함될 가능성이 높다고 판단하였고, 일정한 통제 조건하에서 Piostrisk(2000)의 F-Score와 자산규모에 변동이 되는 유상증자 지표를 추가한 재무건전성지수 중 상관관계가 없거나 작은 변수를 찾고자 하였다. 분석 결과, 1997년 이후 유상증자를 시도한 상장 제조업체의 수는 점차 감소하는 추세로 나타나 유상증자가 재무건전성지수와 지속적인 상관관계가 있지 않음을 밝혔다. 이후 유

상증자를 지수 구성지표에서 제거하고, 신용등급 이전의 연구를 종합하여 0.01 유의수준에서 유의한 자산변동규모 변수를 추가하여 수정 재무건전성지수를 제시하였다(김형민, 2003).

박경덕, 한길석, 윤석진(2008)은 재무비율을 재무건전성지수로 점수화 한 뒤 신용등급과의 관련성을 분석하였다. 분석 결과, 증권거래소와 코스닥 기업의 재무건전성지수가 높은 기업이 신용등급이 높은 것으로 나타나 재무건전성지수와 신용등급 간 밀접한 관계가 있음을 확인하였다. 박경덕, 한길석, 윤석진(2008)은 Piotroski가 제시한 F-Score 모델 일부를 수정한 재무건전성지수를 제시하였다. Piotroski가 사용한 9가지 재무지표에 이자보상배율을 추가하고, 매출액총이익률의 변화는 매출액영업이익률로 바꾸었다. 또한 레버리지 비율인 총자산 대비 장기금융부채 비율을 총자산 대비 총금융부채 비율로 바꾸어 총 10점 만점의 재무건전성지수를 제시하였다.

박경덕, 한길석, 윤석진(2008)은 10개 지표로 구성된 재무건전성지수를 이용하여 상장제조업체 547개 기업의 5개년 재무자료를 분석하였다. 분석 결과, 신용등급이 높은 기업이 재무건전성지수도 높게 나타나 재무건전성지수와 신용등급 간 유의한 상관관계가 있음을 밝혔다. 최상 등급인 AAA의 기업의 경우 분석 기간 동안 재무건전성지수가 7이상을 유지한 것으로 나타났다. 재무건전성이 악화된 상장기업의 수가 증가했던 2000년을 제외하고 등급이 BBB 이상 기업은 재무건전성지수가 6이상으로 나타났고, 등급이 BB 이하 기업은 1999년을 제외하고 6이하로 나타났다. 이러한 분석 결과는 신용등급 결정을 계량화한 정보만 가지고 설명하는 형태로 다소 자의성이 내포될 가능성이 있다. 하지만 이러한 결과는 단순화한 재무건전성지수만으로도 대략적인 신용등급 예측이 가능함을 시사한다.

이장희와 이종열(2013)은 기존에 사용하던 복잡한 수리적 기법의 측정모델 아닌 단순화한 재무건전성지수 모델을 이용하여 회계이익의 질이 미치는 효과를 검증하였다. 또한 재무건전성지수의 개별 재무지표를 통해 회계이익의 질적 속성과 유의성이 높은 새로운 지수모델을 제시할 수 있음을 시사한다.

정준수(2017)는 일반기업에만 적용하던 재무건전성지수를 공공기관에 적용한 연구를 진행하였다. 정준수(2017)는 공공기관이 한국채택국제회계기준

(K-IFRS)을 도입한 첫 해인 2011년회계연도 공공기관 재무제표 자료를 가지고 재무건전성지수를 도출하였다.

정준수(2017)는 Piotroski(2000)가 제시한 F-Score 9가지 변수에 고덕필(2003), 박경덕, 한길석, 윤석진(2008)의 연구를 종합하여 이자보상배율과 자산규모의변화 2가지 변수를 추가하여 총 11개 재무지표로 구성된 재무건전성지수를 도출하였다. 정준수(2017)는 연구를 통해 도출한 재무건전성지수와 공공기관 경영평가 항목이 경영효율 지표와의 유의한 상관관계가 있음을 증명하였으며, 단순화한 지수로 공공기관의 재무적 성과를 파악할 수 있는 지표로서의 활용 가능성을 검증하였다.

정부는 공공기관 회계투명성 제고 및 선진화된 관리기반 구축을 목적으로 2011년 한국채택국제회계기준(K-IFRS) 도입하였다. 이에 따라 공기업은 2011년, 준정부기관은 2013년 회계연도부터 적용하기로 하였다. 이로써 공공기관도 상장기업과 동일한 회계기준을 적용, 원칙적으로 재무 성과의 객관적 비교가 가능하게 되었다(정준수, 2017) 한국채택국제회계기준(K-IFRS) 도입으로 변경된 회계기준 따른 기관 결산 업무와 재무제표 작성이 이루어졌으며, 도입 후 일부 기관은 재무건전성이 제고되는 변화를 보였다(표영호, 2012).

이에 본 연구에서는 공기업 재무건전성 예측모델을 설계하기 위해 공공기관이 한국채택 국제회계기준(K-IFRS)을 도입한 2011년 이후 재무제표 자료를 활용하기로 하였다. 이에 2011년부터 2017년까지 7개년 회계연도 공기업 재무제표 자료를 가지고 먼저 재무건전성지수를 도출하고, 이를 목표변수로 설정하여 4가지 머신러닝 기법을 이용하여 공기업 재무건전성 예측모델을 설계하였다. 공기업 재무건전성지수의 도출 과정은 제 4장 실증분석을 통해 제시하였다.

2.2 머신러닝(Machine Learning)에 관한 연구

2.2.1 머신러닝이란

본 장에서는 머신러닝의 개념과 특징, 머신러닝의 간략한 역사와 머신러닝의 분류 형태에 대해 알아보고, 이어 공공기관과 기업 관련 머신러닝 연구 및 활용 현황, 그리고 본 연구의 분석에서 사용하는 4가지 머신러닝 기법에 대해 살펴보고자 한다.

2.2.1.1 머신러닝의 개념과 특징

최초의 머신러닝 격인 체커스(Checkers) 개발자인 Arthur Samuel(1959)은 머신러닝(Machine Learning)을 '직접적으로 프로그래밍하지 않아도 컴퓨터 혹은 기계가 스스로 학습할 수 있는 능력을 부여하는 연구분야'라고 정의하였다.

하지만 1959년의 정의에서는 데이터의 중요성이 별로 부각되지 않았다. 이에 김승연과 정용주(2017)는 머신러닝은 "데이터를 이용해서 명시적으로 정의되지 않은 패턴을 컴퓨터로 학습하여 결과를 만들어 내는 학문 분야"라고 정의하였다. 즉, 머신러닝은 데이터, 패턴인식, 컴퓨터를 이용한 학습, 이 세 요소가 합쳐져서 만들어진 분야라는 것이다.

최근의 머신러닝 개념은 빅데이터(Big Data), 클라우드 컴퓨팅 등의 환경을 포함해서 이해해야 한다. 나아가 머신러닝은 다양한 확률과 조합 이론, 수학적 최적화 기법, 통계, 알고리즘, 컴퓨터 구조를 활용하여 이상적인 학습 및 예측모델을 구축하는 기술로 연구자의 경험적 지식 습득과 그 응용방법까지 포함하는 융합기술로 발전하고 있다. 즉, 시대의 흐름에 따라 머신러닝의 개념이 재해석 되고 있다(이근영, 2015)

오늘날 머신러닝의 의미는 좀 더 확대된 형태로 사용되고 있다. 머신러닝을 빅데이터 분석 기법의 한 형태로 자동으로 인식된 데이터 패턴을 이용하여 미래를 예측을 하거나 의사 결정을 내리는 하나의 방법으로 정의되고 있다(머피, 2015).

빅데이터란 기존의 소프트웨어는 처리할 수 없을 정도로 크고 복잡한 데이터를 의미한다. 하지만 오늘날 빅데이터는 대량의 데이터를 다루는 문제를 넘어서는 개념으로, 데이터 자체뿐만 아니라 데이터가 제공하는 기회를 활용하기 위한 기술 동향 진보를 의미한다(이형탁, 2019)

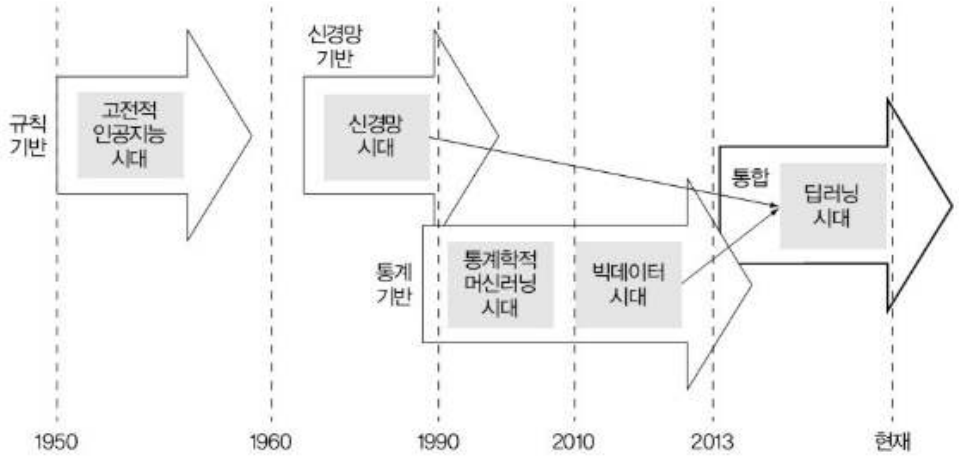
머신러닝(Maching Learing)의 본질은 학습하는 기계를 통한 패턴인식의 자동화된 프로세스이다. 머신러닝의 주요 목적은 많은 복잡한 작업이나 문제를 처리할 때 인간 수준의 능력을 달성하거나 능가할 수 있는 시스템을 구축하는 것이다. 때문에 머신러닝이 빅데이터를 기반으로 한 인공지능(Artificial Intelligence, AI)의 일부라고 할 수 있다. 즉, 컴퓨터 또는 기계로 하여금 복잡한 데이터 또는 문제를 입력하여 스스로 학습하게 하여 인간 수준 능력을 발휘하도록 하는 데 있기 때문이다(Wu, Buyya & Ramamohanarao, 2016).

2.2.1.2 머신러닝의 변천사

머신러닝(Machine Learning)의 시작은 인공지능(Artificial Intelligence, AI)이라고 할 수 있다. 둘 다 컴퓨터를 이용하여 지능적인 작업을 하는 방향으로 발전해왔기 때문이다. 이에 머신러닝을 인공지능의 한 분야로 볼 수 있다. 머신러닝은 1950년대 인공지능이라는 개념으로 태동했고, 신경망 시대를 거쳐 통계학적 머신러닝과 빅데이터 시대를 지나 지금의 딥러닝 시대에 다다랐다. 각 시대별 특징을 보면 다음과 같다(김승연, 정용주, 2017).

1950년대는 고전적 인공지능 시대로 1946년 컴퓨터의 발명 이래 컴퓨터의 가능성에 대한 다양한 논의가 있었다. 특히 유명한 컴퓨터 과학자였던 Alan Turing이 인공지능을 판별하는 튜링 테스트를 제안하였다. 이는 기계가 인간과 얼마나 비슷하게 대화하는지를 기준으로 기계의 지능을 판별하는 데스트였다.

1957년 퍼셉트론(perceptron)이라는 신경망이 개발되었다. 하지만 당시에는 데이터가 굉장히 한정적이어서 신경망 성능이 신통치 않았고, 기초 이론의 부족으로 한정적 패턴만 학습이 가능했다. 여러 학자들이 신경망 구축에 뛰어들었지만, 예상만큼 결과가 나오지 않아 침체 상태가 이어졌다. 이에 신경망 시대를 '인공지능의 겨울'이라고 부르곤 한다.



[그림 2-1] 머신러닝 변천사

출처: 김승연, 정용주(2017)

1990년대 들어 통계학을 전산학과 접목하여 대규모 데이터에서 패턴을 찾는 시도가 있었고, 데이터에 더 큰 비중을 둔 결과 기진일보한 성과가 있었다. 이후 여러 산업 분야에 도입되었는데, 이 시기에 머신러닝이라는 용어가 등장했다. 딥러닝 등장 이후 통계학 중심 기법을 통계학적 머신러닝이라 한다.

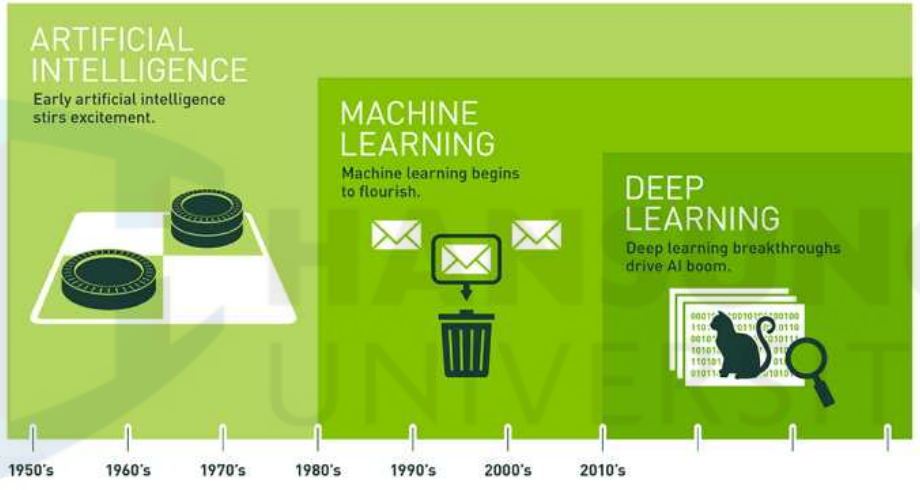
통계학적 머신러닝은 웹을 통해 쏟아지는 테어너와 대용량 저장창지, 분산처리기술과 결합하여 엄청난 시너지를 만들어냈다. 빅데이터(Big Data)라는 용어가 2010년대 유행하였는데, 빅데이터 시대에는 더 큰 데이터 분석을 위한 더 큰 규모의 머신러닝 기법이 만들어지고, 성능 또한 더 좋아지게 되었다.

데이터의 증가와 GPU의 발달로 인한 연산 능력의 증가하면서 머신러닝 연구자들은 신경망 이론에 다시 한 번 눈길을 돌렸다. 그 결과 기존의 신경망보다 훨씬 더 복잡하고 깊이가 있는 신경망을 사용하게 되면서, 이를 딥러닝이라 부르게 된다. 딥러닝 시대는 데이터의 폭발적 증가로 복잡한 신경망을 학습할 수 있는 기반이 생겼다.

현재 머신러닝은 대량의 데이터를 바탕으로 하는 딥러닝 기법을 주로 사용한다. 이를 통해 기존에 해결하지 못하였던 음성 인식, 번역, 이미지 인식에서 좋은 성과를 보이고 있다.

딥러닝(Deep Learing)은 머신러닝의 한 분야로, 일반적인 머신러닝보다 더 발전된 형태를 의미한다. 머신러닝이 데이터를 통해 컴퓨터나 기계에 학습능력을 부여하는 형태라면, 딥러닝은 커퓨터 또는 기계가 학습조차도 스스로 판단하여 실행하고 미래 상황을 예측한다. 즉, 데이터 자체를 신경망 구조를 통해 학습한다. 딥러닝의 대표적인 예가 바로 2016년 3월 전 세계가 주목한 알파고(AlphaGo)와 이세돌 9단의 대국이다(전해남, 2018)

인공지능(Artificial Intelligence), 머신러닝(Machine Learing), 딥러닝(Deep Learnig)의 관계도를 도식화하면 [그림 2-2]와 같다.

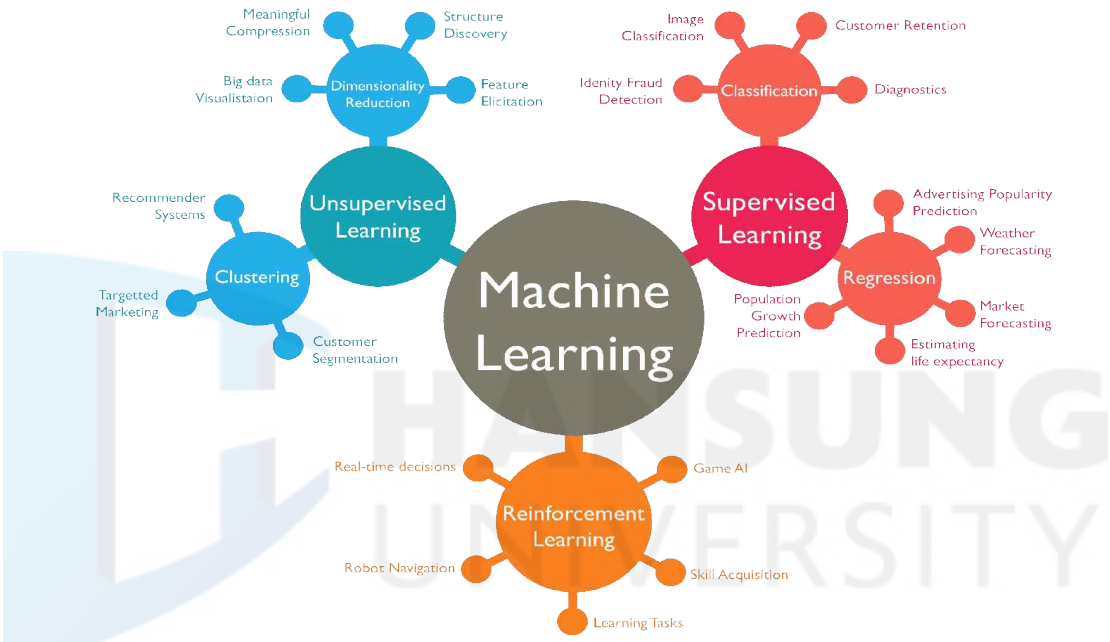


[그림 2-2] 인공지능, 머신러닝, 딥러닝의 관계 다이어그램

출처: https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/

2.2.1.3 머신러닝의 분류

머신러닝(Machine Learning) 기법은 크게 지도학습(Supervised Learning), 비지도학습(Unsupervised Learning), 강화학습(Reinforcement Learning), 이 세 가지로 주로 분류한다(오렐리앙 제롱, 2018)



[그림 2-3] 머신러닝 기법의 3가지 분류

출처: Abdul Rahid (2017)

In https://www.wordstream.com/blog/ws/2017/07/28/machine-learning-applications

지도학습(supervised learning)은 컴퓨터에 데이터(data)와 데이터에 대한 레이블(label), 즉 명시적 정답을 주고 훈련을 시켜서 미래의 출력값을 예측할 수 있도록 하는 학습 방법이다(박유선, 2019). 데이터(data)와 데이터에 대한 레이블(label)을 주고 미래의 출력값을 예측할 수 있도록 하는 학습 방법이다. 지도학습은 입력된 데이터의 특성과 예측 결과값의 연관 관계를 가지고 예측 모델링을 구축한다. 예측값의 특성에 따라 크게 회귀(Regression), 분류(Classification) 두 가지로 나눌 수 있으며, 회귀의 일종으로 보는 추천

(Recommendation)과 랭킹(Ranking)을 추가하여 구분하기도 한다. 예를 들면 이전 주식 시장 변화를 보고 내일 주식 시장 변화를 예측하는 일, 스팸 메일(spam mail), 즉 원치 않는데도 일방적으로 보내어지는 광고성 이메일을 구분하여 분류하는 일, 사용자가 구매한 상품을 토대로 추후 구입 상품을 예측하여 추천하는 일 등이 이에 해당한다. 추천 시스템이나 랭킹의 경우 회귀의 일종으로 볼 수 있으나, 추천시템의 경우 입력과 출력값이 아니라 다양한 관계를 고려하며, 랭킹은 출력값의 예측이 아닌 데이터의 순위를 예측하는 형태이다. 지도학습에 주로 쓰이는 모델로는 선형 회귀(Linear Regression), 로지스틱 회귀(Logistic Regression), 소프트 벡터 머신(Soft Vector Machine, SVM), 결정 트리(Decision Tree), 랜덤 포레스트(Random Foreset, RF), 신경망(Neural Network, NN) 등이 있다. 대부분의 머신러닝의 문제는 지도학습에 해당하는 경우가 많다(김승연, 정용주, 2017).

비지도 학습(unsupervised learning)은 컴퓨터에 데이터(data)를 주되 데이터에 대한 레이블(label)-명시적 정답-을 주지 않은 상태에서 학습하는 유형이다(박유선, 2019). 레이블(label)이 없기 때문에 입력된 데이터로만으로 데이터의 숨겨진 패턴(pattern)이나 구조를 찾아내는 학습 형태이다. 비지도학습은 데이터만 가지고 직접 모델링을 구축하는데, 대표적인 예로 군집화(clustering)와 차원 축소가 있으며, 토픽 모델링(Topic modeing)과 밀도 추정도 있다. 예를 들면 신문기사를 수집하여 비슷한 내용의 기사들을 묶어 각각 경제, 스포츠, 연예 군집으로 나눌 수 있다. 토픽 모델링의 경우 군집화와 유사하지만 주로 텍스트 데이터에 사용되며, 신문 기사를 토픽 모델링한다면 경제 0,1, 스포츠 0.4, 연예 0.5 등 관련 정도를 확률로 표현한다. 각국 학생들의 키와 몸무게를 모아놓은 통계 자료에서 키와 몸무게의 관계를 파악하는 방법이 밀도 추정이다. 차원 축소는 데이터가 복잡하고 고차원인 경우 시각화(visualization)하기 어려울 때 데이터의 잘 나타내는 특징을 추려내기 위해 2차원이나 3차원으로 표현하기 위해 사용한다. 비지도학습에 사용되는 대표적인 기법으로는 K-평균 군집화(K-means Clustering), 커널 밀도 추정(Kernel Density Estimation, KDE), 가우시안 혼합 모델(Gaussian Mixture Model, GMM), 주성분 분석(Principal Component Analysis, PCA) 등이 있다(김승연, 정용주, 2017).

강화학습(reinforcement learning)은 컴퓨터(에이전트)가 주어진 환경과의 상호작용을 통해 장기적으로 얻는 이익이 최대가 되도록 학습하는 형태이다. 즉, 에이전트가 주어진 환경에 대해여 어떤 행동을 취하면 그에 따른 보상을 얻는 시스템을 통해 학습을 진행하되, 에이전트가 보상을 최대화하는 방향으로 행동 또는 행동 순서를 선택한다(김승연, 정용주, 2017; 박유선, 2019). 예를 들어 바둑 프로그램에서 현재 판에서 다음 수의 선택을 학습하는 형태가 지도학습이라면, 강화학습은 게임의 전체 수를 고려하여 게임에서 이길 경우 점수를 받고 그렇지 않은 경우 점수를 주지 않아 일련의 이기는 수를 학습하는 형태이다. 또다른 예로 화성탐사로봇의 경로 탐색과정을 들 수 있다. 지구에서는 화성에 대한 데이터를 얻을 수 없어, 로봇은 실제 환경에 대한 정보가 없다. 따라서 매번 행동에 따른 피드백(보상)을 통해 다음 행동을 정하는 과정을 통해 학습을 진행한다(김승연, 정용주, 2017)

2.2.2 머신러닝 기법을 이용한 공공기관 및 기업 관련 예측 연구

앞서 언급한 세 가지 머신러닝 기법 중 특히 지도학습 기법이 미래에 대한 예측과 예측을 통한 예방 및 대비가 가능하기 때문에 현재 각 산업분야에서는 머신러닝 지도학습 기법을 이용한 연구가 많이 진행되고 있다(이형탁, 2019).

이미 기업에서는 머신러닝 기법을 이용하여 주가 예측이나 부도 예측 등의 연구가 계속 이어져 왔으며, 최근 계속해서 발전하고 있는 다양한 머신러닝 기법을 이용하여 신용평점 예측 등의 연구도 이루어지고 있다. 최근 많이 사용되고 있는 머신러닝 기법을 활용한 기업 관련 연구 내용들을 요약하면 〈표 2-1〉과 같다.

〈표 2-8〉 머신러닝 기법을 이용한 기업 관련 예측 연구(요약)

연구자 (연도)	연구내용	핵심 키워드	연구자료	분석방법	연구결과
김성진, (2015)	“랜덤 포레스트를 활용한 기업채권등급평가 모형”	랜덤 포레스트, 채권등급평가, 다분류 분석	KOSPI 또는 KOSDAQ 상장 1295 제조업 데이터	RF, 다변량판별분석(MDA), 인공신경망, MSVM	신용등급 평가시 랜덤 포레스트 방법이 전통기법보다 빠르고 정확한 예측 결과를 산출함
양진용 (2017)	“기업 재무 정보 활용 머신러닝 기반 경영 예측 시스템”	머신러닝, 예측 시스템, 기업 경영, 파산 예측, 주가 예측	건설 기업	머신러닝 분석 – AdaBoost, , DNN, SVM, D.Tree	AdaBoosting 기법의 예측 결과 우수
최정원, 오세경, 장재원 (2017)	“빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구”	기업부도예측, 텍스트마이닝, Word2vec, 인공지능, 머신러닝	2001~2015년 상장 기업 부도기업, 2010-2016 해당 부도 기업 기사	Lgit, DT, RF, SVM, DNN, RNN, KMV	RF 예측력이 가장 우수. 재무 정보가 불투명한 중소기업의 경우, 인공신경망 예측 모형이 더욱 효과적임을 밝힘
이현상 오세환 (2019)	“머신러닝 기법을 활용한 기업 신용 평점 예측모델 개발”	KIS, 신용평점, 회사채 신용등급, 머신러닝, 딥러닝, 예측모델	2,337개기업자료 (KISVALUE 자료)	SVM, MLP,RF, DNN	Random Forest가 가장 예측력이 높음

최근 공공 데이터를 가지고 머신러닝 방법을 이용한 많은 예측 연구들이 이루어지고 있다. 산업재해 예측(최종고, 2019), 교통수요 예측(김현주, 2019) 등이 이에 해당한다. 하지만, 머신러닝 기법을 이용한 공공기관의 경영성과나

재무적 성과 예측 등의 연구는 찾아보기 어려운 상황이다. 기업 사례와 마찬가지로 머신러닝 기법을 활용한 공공기관 관련 연구 내용을 요약하면 〈표 2-9〉와 같다.

〈표 2-9〉 머신러닝 기법을 이용한 공공기관 관련 예측 연구(요약)

연구자 (연도)	연구내용	핵심 키워드	연구자료	분석방법	연구결과
하지은 (2017)	"RandomForest와 XGBoost를 활용한 민원 카테고리 및 담당부서 자동분류 성능 비교"	공공데이터, 자동분류, open data, civil petition, RandomForest, XGBoost, classification	서울시 응답소 2010-2017년 8개년 민원 데이터	Random Forest, XGboost	XGBoost가 전반적으로 높게 나타남. 민원 처리업무 프로세스 효율 및 처리시간 단축가능성 확인.
김현주 (2019)	"빅데이터 및 인공지능을 활용한 지방지역 교통수요예측"	인공지능, 지방도로, 교통수요예측, CBR, 인공신경망	통계청, 지자체, 국토교통부 시스템 교통 관련 데이터	빅데이터 활용 CBR, 인공신경망	인공지능을 활용한 교통수요예측 가능성 확인
최종고 (2019)	머신러닝을 활용한 건설근로자 중대사고 위험성 예측모델	산업재해, 건설안전, 랜덤포레스트, 로지스틱 회귀분석, 에이다부스트	2011-2016년 산업재해자 데이터	머신러닝 RF, LR, Adaboost	렌덤포레스트의 AUROC가 가장 예측률(92.62%) 높음. 사고 발생시 사망 위험성 높은 근로자 분류 가능

최근 2-3년 사이 국내 다양한 학문 분야에서 머신러닝 방법을 이용한 연구가 급증하고 있다. 이에 본 연구에서는 공기업 재무자료를 가지고 머신러닝 기법을 이용하여 공기업 예측모델을 설계하여 도출된 예측값과 실제값을 비교한 실증연구를 진행해 보고자 한다. 또한 예측모델을 통해 도출된 요인의 특성을 분석하여 향후 일반 기업이나 공공기관 컨설팅 관련 인사이트를 제공하고자 한다.

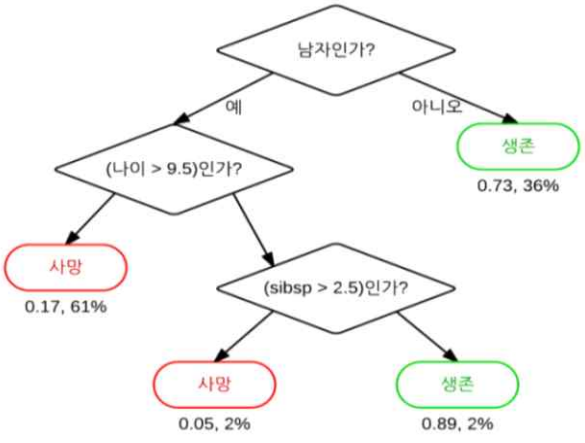
2.2.3 머신러닝 예측모델 이론

본 절에서는 머신러닝 기법 중 지도학습(Supervised Learning)의 형태인 Random Forset, XGboost(Extreme Gradient Boosting), LightGBM(Light Gradient Boosting Machine)과 DNN(Deep Neural Network, 심층신경망), 이 네 가지 기법을 이용하여 공기업 재무건전성 예측모델을 만들고자 한다. 이에 Random Forset, XGboost, LightGBM, DNN에 대한 개념 및 주요 특징들을 차례로 소개하기로 한다.

2.2.3.1 Random Forest

Random Forest는 의사결정트리(Decision Tree) 기반의 알고리즘으로, 다른 성능을 갖는 여러 개의 의사결정트리를 결합하여 만들어낸 앙상블(ensemble) 기법을 말한다(Breiman, 2001).

의사결정트리(Decision Tree)는 데이터를 일정한 규칙에 의해 분류하여 예측하는 기법으로, 그 모양이 트리와 같다고 하여 의사결정트리라 불리운다. [그림 2-4]는 타이타닉호 탑승객 생존여부를 나타내는 의사결정트리이다.

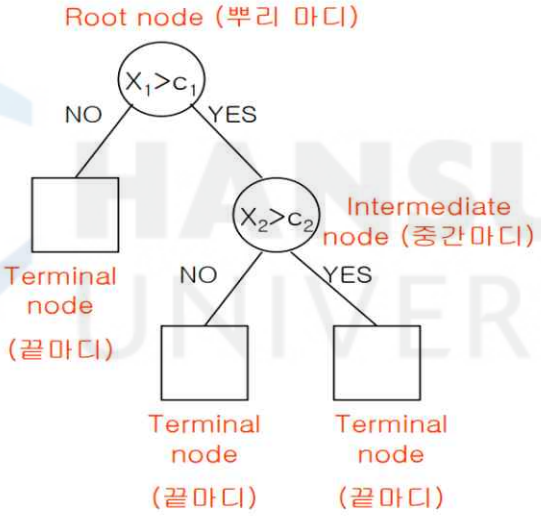


[그림 2-4] 타이타의닉호 탑승객 생존여부를 나타내는 의사결정트리

출처: https://ko.wikipedia.org/wiki/결정_트리_학습법

[그림 2-4]에서 “sibsp”는 탑승한 배우자와 자녀의 수를 의미한다. 여성인 경우 생존 가능성이 높았으며, 남자이면서 나이가 9.5세 이상인 경우 사망 가능성이 높았다. 타이타닉 생존자 예측모델 만들기는 캐글(Kaggle) 튜토리얼(Tutorial)용 경진대회를 통해 소개되어 대표적인 예로 많이 활용되고 있다.

의사결정트리를 일반화하면 [그림 2-5]와 같으며, 뿌리노드(Root node), 중간노드(Intermediate node), 터미널노드(Terminal node)로 이루어져 있다. 터미널노드를 리프(leaf)노드라고도 부른다. 의사결정트리는 각 터끝마디에 속하는 데이터의 개수를 더하면 뿌리노드 데이터 개수와 일치한다. 즉, 끝마디 간 서로 교집합을 가지지 않는다(김재휘 & 김재희, 2019).



[그림 2-5] 의사결정트리의 구성요소

출처: https://ratsgo.github.io/machine%20learning/2017/03/26/tree/

의사결정트리는 목표변수가 어떤 형태의 데이터인지에 따라 크게 분류 트리(Classification Tree)와 회귀 트리(Regression Tree)로 분류한다. 분류 및 회귀 트리를 아울러 CART(Classification And Regression Tree)라고 부르며, 이는 Breiman에 의해 처음 사용되었다[4].

4) 출처: https://ko.wikipedia.org/wiki/결정_트리_학습법

분류 트리와 회귀 트리(CART; Classification And Regression Tree)는 일정 부분 유사하지만, 분할하는 과정에는 차이가 있다. 분류 트리는 가장 빈도가 높은 범주에 새로운 데이터를 분류하며, 회귀 트리는 목표변수의 평균을 예측값으로 반환한다(김재휘, 김재희, 2019).

의사결정트리는(Decision Tree)는 이진 트리를 생성하여 적합한 기준으로 분할할 최상의 범주 또는 수치 변수를 찾는다. 분류(Classification)에서는 지니 불순도(Gini impurity) 또는 엔트로피(Entropy)를 사용하며, 회귀(Regression)에서는 평균제곱오차인 MSE(Mean Square Error)나 평균절대오차인 MAE(Mean Absolute Error)를 사용하여 분산(Variance) 감소를 계산한다[5].

분류(Classification)에서 지니 불순도(Gini impurity)와 엔트로피(Entropy)를 계산하는 수식은 (2-1), (2-2)와

$$I_G(f) = \sum_{i=1}^{n} f_i(1-f_i) = 1 - \sum_{i=1}^{n} {f_i}^2 \quad (2\text{-}1)$$

$$I_E(f) = \sum_{i=1}^{n} f_i \log_2(f_i) \quad (2\text{-}2)$$

각 수식에서 i는 i번째 마디(node), n은 결과값 클래스의 개수를 의미한다.

회귀(Regression)에서는 평균제곱오차인 MSE(Mean Square Error)와 평균절대오차인 MAE(Mean Absolute Error)의 수식은 (2-3), (2-4)와 같다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2 \quad (2\text{-}3)$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y_i} - y_i\right| \quad (2\text{-}4)$$

각 수식에서 n은 데이터 수, $\hat{y_i}$는 예측값, y_i는 실제값을 의미한다.

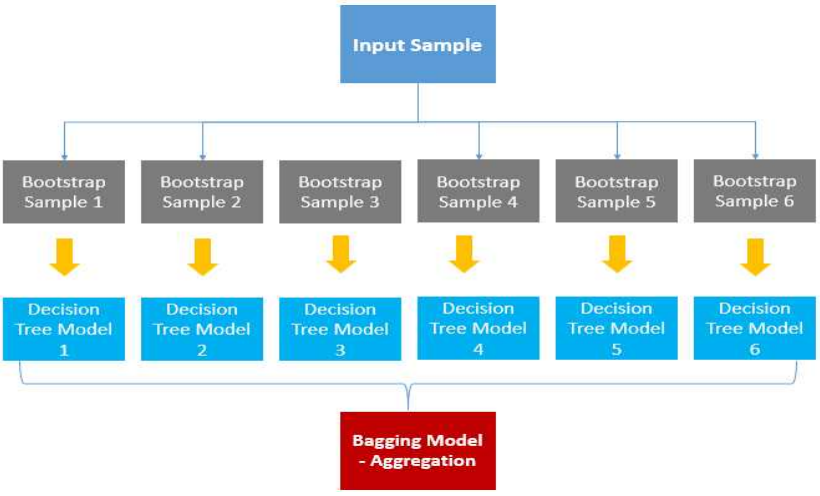
의사결정트리(Decision Tree)는 데이터를 분할하는데 있어서 최상의 방법

5) 출처: https://scikit-learn.org/stable/modules/tree.html#tree-algorithms-id3-c4-5-c5-0-and-cart, https://ko.wikipedia.org/wiki/결정_트리_학습법

을 찾기까지 가지치기를 통해 완성한다. 또한 데이터가 나뉘어져 갈라지는 분기 시 하나의 변수만을 사용하기 때문에 결과가 명확하여 설명이 용이하다. 하지만 반대로 분기 시 하나의 변수만을 사용하기 때문에 비슷한 설명력을 가진 다른 변수들은 모델 구축 시 고려하지 않고, 오로지 학습 데이터만을 잘 분류하는 분기 변수들로만 모델이 구축된다. 학습 데이터를 따라 모델 구성이 크게 달라지는 등 모델의 불안정성 문제를 야기하게 되고, 트리 형태의 일반화가 어려워 과적합(over-fiting) 문제가 발생하게 된다. 결과적으로 예측력이 떨어지며, 모델의 안정성 또한 떨어진다(Li & Belford, 2002; 하지은, 2019; 이형탁, 2019)

앙상블(ensemble)은 머신러닝 분야에서는 여러 개의 모델들을 활용해 종합하여 예측하는 기법을 말한다. 적당한 예측력을 가진 모델 하나 보다는 여러 모델을 종합하여 결정하면 더 나은 의사결정을 할 수 있음을 의미한다(김동영, 원동은, 2019). 앙상블 기법의 대표적인 예로 배깅(Bagging)과 부스팅(Boosting)을 들 수 있다. 배깅이란 Bootstrap Aggregation의 줄임말이다. 배깅은 [그림 2-6]처럼 주어진 데이터로 여러 개의 부트스트랩(boostrap)을 생성하여 각 부트스트랩 별로 모델링을 한 후 결과를 집계(aggregating)하여 최종 예측모델을 산출해내는 기법으로, 병렬 형태로 학습시키는 형태이다. 여기서 부트스트랩이란 단순 복원 임의 추출법을 통해 기존의 데이터로부터 추출한 크기가 동일한 여러 개의 표본자료를 말한다(이형탁, 2019; 하지은, 2019).

의사결정트리는 대체로 분산이 높다는 문제가 있다. 심지어 훈련 데이터를 임의로 두 집단으로 나눠 각각 의사결정트리 모델을 적용했을때 두 모델의 결과는 매우 다를 수 있다. 이를 보완하기 위해 부트스트랩(Bootstrap)을 적용한 기법이 바로 배깅(Bagging)이다. 배깅은 하나의 훈련 데이터로부터 여러 개의 부트스트랩 샘플을 만들어 각 샘플을 다른 의사결정트리에 적용한다. 이렇게 나온 각 의사결정트리의 결과들을 모아 평균을 구함으로써 분산을 줄일 수 있다(김재휘, 김재희, 2019).

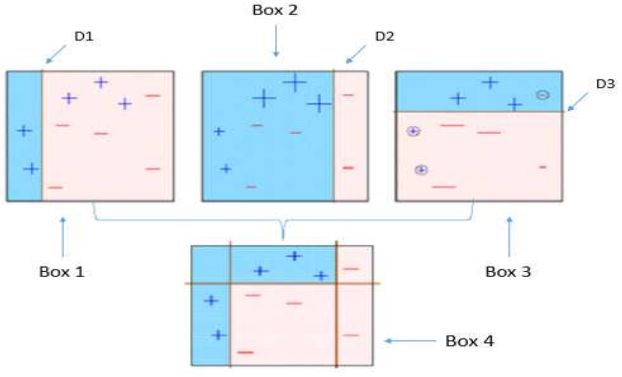


[그림 2-6] 배깅(Bagging) 기법 흐름도

출처: https://swalloow.github.io/bagging-boosting

배깅(Bagging)은 학습데이터의 작은 변화가 예측 결과의 큰 변화를 야기시키는 불안정한 데이터에 효과적이지만, 데이터가 안정적일 경우에는 성능의 향상을 기대하기 어렵다(Breiman, 1996; Opitz, Maclin, 1999).

부스팅(Boosting)은 약한 학습 데이터를 강한 학습 데이터로 변환시켜 오차를 교정하는 과정을 반복하면서 예측모델의 정확도 향상시키는 기법으로, 순차적으로 학습을 시키는 형태를 취한다(이형탁, 2019; 하지은, 2019). [그림 2-7]은 부스팅(Boosting)의 한 형태인 Adaboost의 예이다.



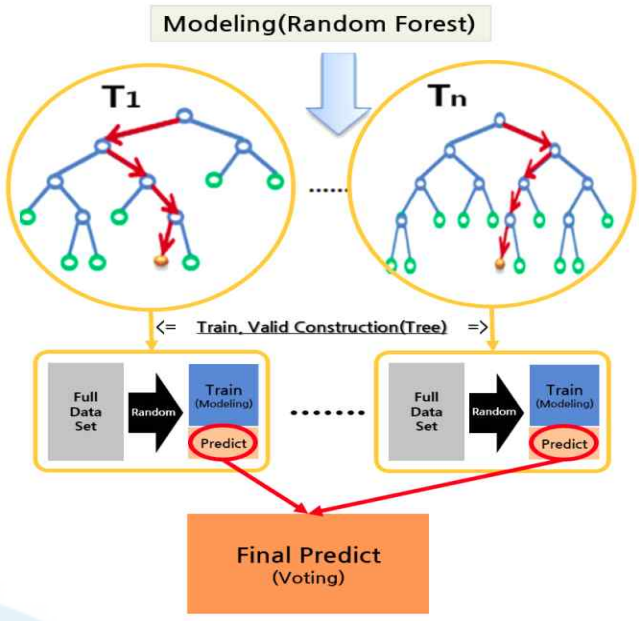
[그림 2-7] 부스팅(Boosting)의 한 형태 : Adaboost 기법

출처: https://medium.com/greyatom/a-quick-guide-to-boosting-in-ml-acf7c1585cb5

각 Box는 10개의 데이터 포인트로 플러스(+) 5개와 마이너스(-) 5개로 구성되어 있으며, 동일한 가중치를 부여받는다. Box1 모델에서 D1은 세로 선을 분리하여 만들었지만, 적색 영역에 플러스(+) 3개를 잘못 분류하였다. Box2 모델에서는 잘못 분류된 3개의 플러스(+)에 다른 관측 포인트보다 더 많은 가중치가 부여되며, D2 또한 세로선을 분리하여 만들었지만, 두 개의 마이너스(-)를 잘못 분류한다. 이와 같은 다음 모델은 가장 정확한 예측 변수가 작성될 때까지 계속 진행하고 이전 모델이 직면한 오류를 조정한다. 최종적으로 Box4 모델은 Box1, Box2 및 box3 모델을 결합하여 사용된 개별 모델보다 훨씬 우수한 강력한 예측모델을 만든다.

앞서 언급한 의사결정트리(Decision Tree) 모델에 배깅(bagging) 기법을 적용한 앙상블 모델을 Random Forest라고 한다. 즉, 전체 데이터셋에서 임의로 데이터들을 만들어 여러 개의 학습된 의사결정트리 모델을 만들고 이를 종합하는 방식이다(김동영, 원종은, 2019)

Random Forest를 이루고 있는 하나의 트리는 모델이 형성될 때 전체 학습 데이터로부터 독립변수와 개체를 무작위 복원 추출한 일부 데이터셋만을 가지고 학습을 수행해나간다. 이런 개별 의사결정트리는 저마다 다른 결과 값을 가지게 되는데 Random Forest는 각 트리마다 다르게 예측된 값을 모두 모아 평균을 내거나 또는 투표를 통해 최종 예측값을 산출한다(심재헌, 2016; 김태식, 홍정식, 전윤수, 박종율, 안태욱, 2018). Random Forest 모델링 과정을 도식화하면 [그림 2-8]과 같다.



[그림 2-8] Random Forest 모델링 프로세스

Random Forest의 분류(Classification) 트리와 회귀(Regression) 트리의 수식은 (2-5), (2-6)과 같다.

$$\widehat{C}_{rf}^{n}(x) = majority\ vote\ \ \widehat{C}_{b}(x)_{1}^{n} \quad \cdots (2\text{-}5)$$

여기서 $\widehat{C}_{b}(x)$는 b번째 Random Forest 트리를 의미한다.

$$\widehat{f}_{rf}^{n}(x) = \frac{1}{n}\sum_{b=1}^{n} T_{b}(x) \quad \cdots (2\text{-}6)$$

여겨서 $T_{b}(x)$는 Random Forest 트리의 클래스 예측값을 의미한다.

의사결정트리(Decision Tree)에서는 모든 변수를 사용하여 가장 최적의 결과를 도출하는 것과 달리 Random Forest에서는 독립변수를 무작위로 선택하고, 선택된 독립변수의 집합 중에서 가장 최적의 결과를 내는 방법을 이용한다

(권안나, 2013). 이에 따라 Random Forest는 임의성을 최대로 가지게 되어 의사결정트리 간 상관관계를 낮추어 분산을 줄여준다. 때문에 의사결정트리의 수가 많을수록 예측 오차가 줄어 과적합(over-fitting) 현상이 발생하지 않고, 높은 예측력을 보이며 매우 안정적인 모델 형태를 제공한다(권안나, 2013; Siroky, 2009; Breiman & Cutler, 2014). 또한 독립변수의 개수가 수 천개라도 제거하지 않고 모두 활용할 수 있기 때문에 규모가 큰 데이터 자료에서 중요한 변수를 찾는데 용이하다. 특히 독립변수의 개수가 많을 때에는 배깅(bagging)이나 부스팅(boosting)과 비슷하거나 더 나은 예측력을 보이는 경우가 많은 것으로 알려져 있다(권안나, 2013). Random Forest는 목표변수가 이항변수 혹은 다항변수 여부에 관계없이 동일하게 적용될 수 있어 기업의 부실 예측이나 신용 등급 예측에도 활용되고 있으며(김성진, 2015), 최근 공공기관 데이터를 이용한 산업재해 예측에도 활용되고 있다(최종고, 2019)

Random Forest는 변수 중요도(variable impotance)를 제공한다. 일반적으로 변수의 중요도를 plot 함수로 표현하는데, plot은 변수의 점수를 보여준다. 변수의 중요도 점수가 높을수록 예측 변수에 대한 기여도가 높다고 본다. 한편 Random Forest에서 변수의 중요도는 분류(clasification) 트리는 지니(Gini), 회귀(regression) 트리는 MSE(Mean Square Error, 평균제곱오차)로 측정할 수 있다. 분류(clasification) 트리에서는 트리가 작성되면 Gini 불순도를 계산하여 각 노드에서 분할할 변수를 결정한다. 각 변수에 대해 Forest의 모든 트리에서 해당 변수를 선택하여 노드를 분할할 때마다 Gini 감소의 합은 누적된다. 합계는 평균을 제공하기 위해 Forest의 트리 수로 나뉜다. 척도는 관련이 없고, 상대적인 값만 중요하다. 중요도는 두 측정값 사이에 대략적으로 정렬되며 숫자 변수는 Gini 척도에서는 더 낮다. 이는 잠재적으로 분할점이 많기 때문에 숫자 변수를 사용한 노드 분할에 대한 편향을 나타낼 수 있다. 회귀(regression) 트리에서는 MSE(Mean Square Error, 평균제곱오차)의 증가로 계산하며, OOB(Out of bag)의 값을 섞어서 계산한다 노드 순도의 증가는 변수를 나눌 때마다 제곱오차의 감소를 기반으로 계산된다. 한편 두 가지 방법 모두 상관된 예측 변수의 중요성을 과장할 수 있다는 단점이 있다[6].

6) 출처: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1796903/

2.2.3.2 XGboost(eXtreme Gradinet Boosting)

XGBoost는 Chen과 Guestrin(2016)이 소개한 기법으로 선형 모델이나 트리 기반 모델의 과적합(ove-fitting) 문제를 해결하고, 규모가 큰 데이터셋(dataset)의 안정성과 훈련 속도 향상 목적으로 만들어졌다. XGBoost는 eXtreme Gradient Boosting의 약자로, Gradient Boosting 용어는 Friedman의 『Greedy Function Approximation : Gradient Boosting Machine』 논문에서 유래했다.

XGBoost는 회귀 및 분류, 순위 목표를 지원하는 모델로 트리 기반 앙상블 학습에서 각광받는 분석 모델이다. 처리 속도가 빠르며, 모델의 이전 결과를 활용하여 모델을 계속적으로 개선하고 훈련하는 등 성능이 뛰어나 Google, MS Azure, Alibaba 등 실무에서도 많이 활용되었다(하지은, 2017). 또한 예측모델 및 분석 대회 플랫폼인 캐글(kaggle)에서도 많이 활용하였는데, 특히 2015년 한 해 17개 팀들이 XGBoost 기법을 이용하여 대회에서 우승한 기록이 있다(Chen & Guestrin, 2016).

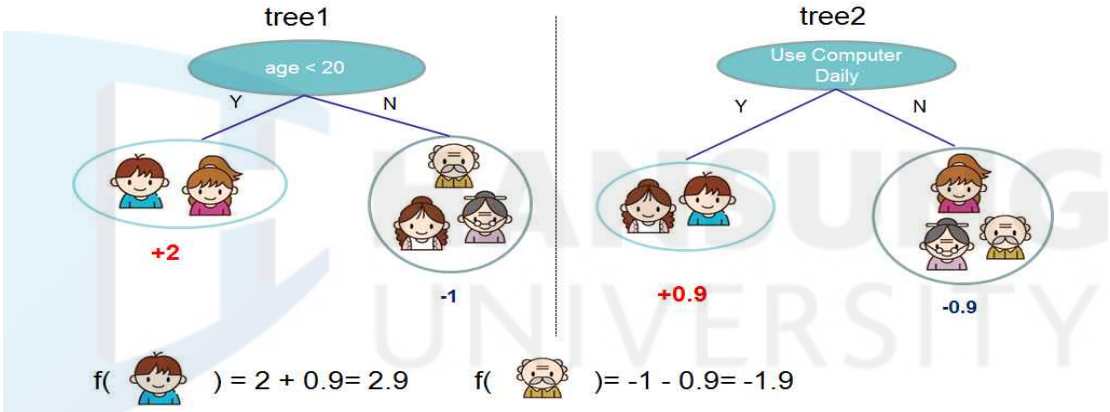
부스팅(Boosting)은 이전 트리로부터 얻은 정보를 가지고 다음 트리를 생성하는데 활용하는데, 약한 개체들을 강한 개체들로 변환시키면서 오차를 교정하는 형태를 취한다. 일반적인 Gradient Boosting에서는 트리 가지치기 과정에서 부정적인 손실(negative loss)이 발생하면 그 과정을 멈춘다. 하지만 XGBoost는 학습 시에 파라미터(parameter)로 지정한 max_depth까지 진행한 후, 손실 함수(loss function)에서 개선이 일정 수준에 미치지 못할 경우 역방향으로 가지치기 과정을 진행한다(Chen & Guestrin, 2016; Friedman, 2001).

이처럼 XGBoost는 결측치를 내부적으로 자동 처리하며 트리를 생성할 때 병렬적으로 생성하며 새로운 데이터에 대해 이전의 학습 모델 결과를 반영하여 그 성능을 더 향상시키는 방법으로 훈련하는 특징을 가진다(Chen & Guestrin, 2016).

Chen과 Cusestrin(2016)에 따르면, XGBoost는 훈련 손실(loss)을 최소화하며 과적합을 줄이기 위해 트리의 복잡도(complexity)를 통제하면서 가장 최적의 모델을 만든다.

XGBoost는 CART(Classification And Regression Tree)라 불리우는 앙상블 모델을 사용하여 트리를 만든다. 이후 부스팅(Boosting)을 통해 각 분류기간 비중(weight)을 최적화한다.

CART 모델은 일반적인 의사결정트리(Decision Tree)와는 조금 다르다. 일반적인 의사결정트리는 잎(leaf) 하나에 대해서만 결정값(decision value)을 갖지만, CART 모델은 모든 잎(leaf)들이 모델의 최종 스코어에 연관되어 있다. 일반적인 의사결정트리는 분류를 제대로 했는지 여부에 대해서만 초점을 맞추는 반면, CART는 같은 분류 결과를 갖는 모델들의 우위를 점수(Score)로 비교할 수 있다.



[그림 2-9] 트리 앙상블 모델의 최종 결정값 계산

출처: Chen(2014)

[그림 2-9]는 “컴퓨터 게임을 좋아하는가?“에 대한 두 가지 트리 모델을 보여주고 있다. 트리 하나에 딸린 모든 잎(leaf)에는 결과값으로 분류된 사람들의 컴퓨터 게임 선호도에 대한 점수가 남는다. 즉, tree1에서 왼쪽 잎 20세 미만 그룹의 남성과 여성은 2점으로, 컴퓨터 게임을 좋아하는 사람이다. 반면 오른쪽 잎의 20세 이상 그룹 3명은 −1점으로 컴퓨터 게임을 선호하지 않는 것으로 보인다. tree2에서 컴퓨터를 매일 사용하는 그룹의 남성과 여성은 0.9점, 그렇지 않은 그룹의 3명은 −0.9점이다. 여기서 tree1과 tree2를 조합하면, 컴퓨터를 게임을 좋아하는가에 대한 각 사람의 최종 점수를 얻을 수 있다.

즉, 트리 앙상블 모델에서는 tree1과 tree2를 조합한 형태로 개별 예측점수(prediction score)를 합산하여 최종 점수를 계산한다. 즉, 앙상블 모델은 트리 간 모델을 상호보완하는 형태다.

Chen(2014)에 따르면, K 개의 트리가 있다고 가정하면, 트리 앙상블에 대한 모델식과 최적화할 목적 함수(Objetive Funtcion)는 (2-7), (2-8)과 같다.

$$\hat{y_i} = \sum_{k=1}^{K} f_k(x_i),\ f_k \in F \qquad \cdots (2\text{-}7)$$

수식 (2-7)에서 $\hat{y_i}$는 예측점수, K는 트리의 개수로, [그림 2-9]는 트리의 개수가 2개인 모델이다. F는 CART로 알려진 회귀 트리들의 공간을 의미한다. f는 F 공간의 함수를 의미한다.

$$Obj(\theta) = \sum_{i=1}^{n} l(y_i, \hat{y_i}) + \sum_{k=1}^{K} \Omega(f_k) \qquad \cdots (2\text{-}8)$$

목적 함수(Objetive Funtcion)는 훈련 손실(Traing Loss)과 트리들의 복잡도(Complexity of the Trees)의 합으로 구성된다. 이제 XGBoost의 트리에 대한 학습을 진행한다. 각 트리별로 트리의 구조와 잎(leaf) 점수를 가진 함수 f가 있다. t번째 최적화할 목적 함수를 (2-9)와 같이 정의해보자.

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \qquad \cdots (2\text{-}9)$$

이제 가지고 있는 트리에 대한 파라미터를 최적화하면 되는데, 기존 최적화(SGD) 함수는 한 번에 여러 개의 트리 최적화가 어렵다. 따라서 기존 모델 함수를 변경하여 한 번에 하나의 트리를 추가하여 각 단계(t)별 결과값을 예측한다. 그 전개 수식은 (2-10) 같다.

$$
\begin{aligned}
\hat{y}_i^{(0)} &= 0 \\
\hat{y}_i^{(1)} &= f_1(x_i) = \hat{y}_i^{(0)} + f_1(x_i) \\
\hat{y}_i^{(2)} &= f_1(x_i) + f_2(x_i) = \hat{y}_i^{(1)} + f_2(x_i) \\
&\cdots \\
\hat{y}_i^{(t)} &= \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)
\end{aligned}
\qquad \cdots (2\text{-}10)
$$

각 단계에서 어떤 트리를 선택해야 할까? 목표는 수식 (2-11)에서 목적 함수를 최소화할 f_t를 찾는 것이다. 각 단계에서 트리의 손실(Loss)를 계산해야 하며, 단계별로 손실 확인이 가능하다.

$$
\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t)}) + \sum_{i=1}^{t} \Omega(f_i) \\
&= \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant
\end{aligned}
\qquad \cdots (2\text{-}11)
$$

만약 손실 함수(Loss Funtion)가 평균제곱오차인 MSE(Mean Square Error)라면, (2-12)와 같은 수식이 도출된다.

$$
\begin{aligned}
Obj^{(t)} &= \sum_{i=1}^{n} (y_i - (\hat{y}_i^{(t-1)} + f_t(x_i)))^2 + \Omega(f_t) + const \\
&= \sum_{i=1}^{n} [2(\hat{y}_i^{(t-1)} - y_i) f_t(x_i) + f_t(x_i)^2] + \Omega(f_t) + const
\end{aligned}
\qquad \cdots (2\text{-}12)
$$

제곱 손실의 경우를 제외하고는 여전히 목적 함수가 복잡해보인다. 이때 Tayler 급수를 이용하여 목적함수의 근사치를 만들면 수식 (2-13)과 같다.

$$
Obj^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) + const \qquad \cdots (2\text{-}13)
$$

여기서 g 와 h 함수는(2-14), (2-15)와 같이 정의한다.

$$g_i = \partial_{\hat{y}^{(t-1)}} l(y_i, y_i^{(t-1)}) \quad \cdots (2-14)$$

$$h_i = \partial^2_{\hat{y}^{(t-1)}} l(y_i, \hat{y}_i^{(t-1)}) \quad \cdots (2-15)$$

기존 공식에서 상수를 제거하면 다음과 같이 미분 형태의 계산만으로 손실 함수를 재정의 할 수 있으며, 새로운 트리에 대한 학습을 최적화할 수 있는 t 단계의 최종 목표 함수식 (2-16)이 도출된다.

$$Obj^{(t)} \simeq \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \quad \cdots (2-16)$$

여기서 함수 g와 h는 t번째보다 작은 형태(t-1 번째)의 미분이므로, 무한히 전개했을 때 수렴하는 형태를 찾을 수 있다. XGBoost는 g와 h 함수만 수정하면, 다양한 손실 함수 지원이 가능하다. 트리 함수 f_x 를 (2-17)과 같이 정의해본다.

$$f_t(x) = w_{q(x)}, \quad w \in R^T, q : R^d \rightarrow 1, 2, \cdots, T \quad \cdots (2-17)$$

여기서 w는 트리의 잎의 가중치, q는 트리의 구조로 입에 할당되는 함수, T는 마디의 개수를 의미한다. 트리 한 개의 복잡도(Complexity of a Tree)를 정의하면 (2-18)과 같다.

$$\Omega(f_t) = \gamma T + \frac{1}{2} \lambda \sum_{j=1}^{T} w_j^2 \quad \cdots (2-18)$$

j번째 잎에 있는 인스턴스 세트 $I_j = i | q(x) = j$로 정의하고, 위의 수식을 이용하여 목적함수를 다시 정리하면 수식 (2-19)와 같다.

$$Obj^{(t)} \simeq \sum_{i=1}^{n}[g_i f_t(x_i) + \frac{1}{2}h_i f_t^2(x_i)] + \Omega(f_t) \quad \cdots (2\text{-}19)$$
$$\simeq \sum_{i=1}^{n}[g_i w_{q(x_i)} + \frac{1}{2}h_i w_{q(x_i)}^2] + \gamma T + \frac{1}{2}\lambda\sum_{j=1}^{T} w_j^2$$
$$\simeq \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \gamma T$$

위의 수식에 $G_j = \sum_{i\in I_j} g_i$, $H_j = \sum_{i\in I_j} h_i$ 를 대입하면 (2-20)과 같이 수식이 더 단순해진다.

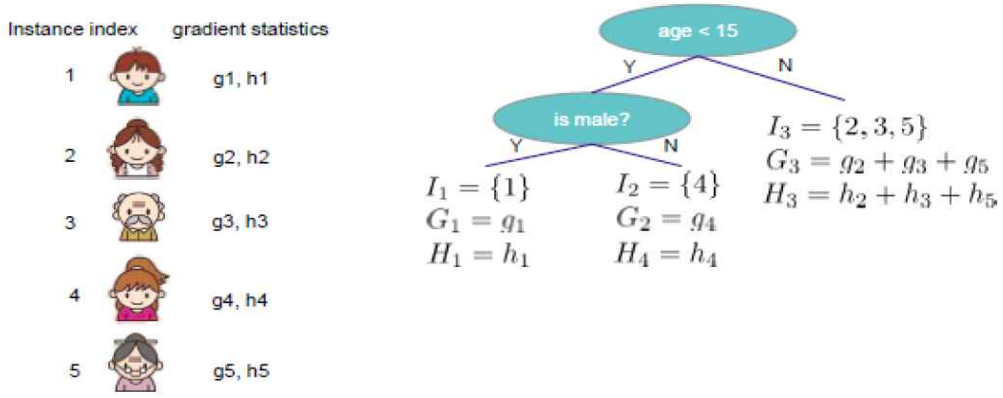
$$Obj^{(t)} = \sum_{j=1}^{T}[(\sum_{i\in I_j} g_i)w_j + \frac{1}{2}(\sum_{i\in I_j} h_i + \lambda)w_j^2] + \gamma T \quad \cdots (2\text{-}20)$$
$$= \sum_{j=1}^{T}[G_i w_j + \frac{1}{2}(H_j + \lambda)w_j^2] + \gamma T$$

여기서 트리의 구조 q(x)는 고정되어 있고, 각 잎(leaf)에 최적의 가중치를 계산하면 w_j^* , 목표 함수는 Ojb^* 이며, 각각의 수식은 (2-21), (2-22)와 같다.

$$w_j^* = -\frac{G_j}{H_j + \lambda} \quad \cdots (2\text{-}21)$$

$$Obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \quad \cdots (2\text{-}22)$$

이 목적값을 통해 모델(트리 구조)이 얼마나 좋은지를 평가하는데, 이 값이 낮을수록 더 좋은 트리 구조를 의미한다.



[그림 2-10] 트리 구조 점수 계산

[그림 2-10]에서 나이(age)가 15세 미만을 기준으로 트리를 분할한 형태이며, 목적 함수값은 수식 (2-23)과 같다.

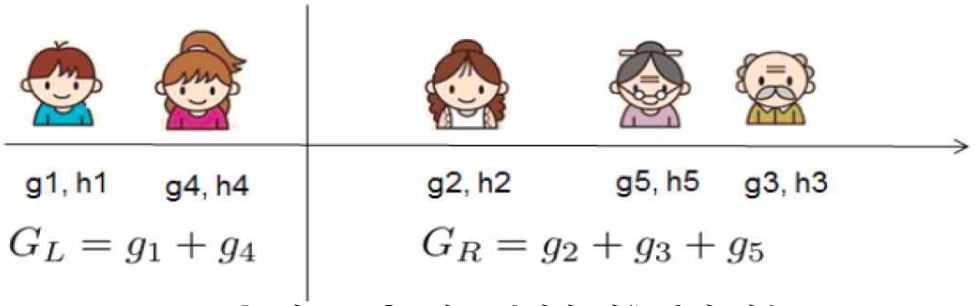
$$Obj = -\sum_j \frac{G_j^2}{H_j + \lambda} + 3\gamma \qquad \cdots (2\text{-}23)$$

이상적으로는 가능한 모든 트리를 생성하고, 이 중에서 가장 최적의 트리를 선택한다. 하지만, 무한한 나무의 구조가 있을 수 있다. 그러므로 실제로 나무를 탐욕럽게(greedily) 성장시킨다. 깊이가 0인 트리에서 시작하여, 하나의 잎을 2개로 분리한다. 분리 후 목적함수의 변화는 수식 (2-24)와 같다.

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \qquad \cdots (2\text{-}24)$$

만약 잎(lear)을 분리한 점수(Gain)가 γ 보다 작면, 잎(leaf)을 분리하지 않는데 좋다. 실제값의 데이터에 대해, 일반적으로 최적의 분할을 찾고자 한다. 이를 효율적으로 수행하기 위해 [그림 2-11]과 같이 모든 인스턴스를 정렬된 순서로 배치한다. 분할 각 잎(leaf)에 대해 모든 변수들을 열거하라. 각 변수에 대해 인스턴스를 변수값 별로 정렬한다. 선형 스캔을 사용하여 해당 기능

에 가장 적합한 분할을 결정한다. 모든 변수에 따라 최상의 분할 솔루션을 취한다.

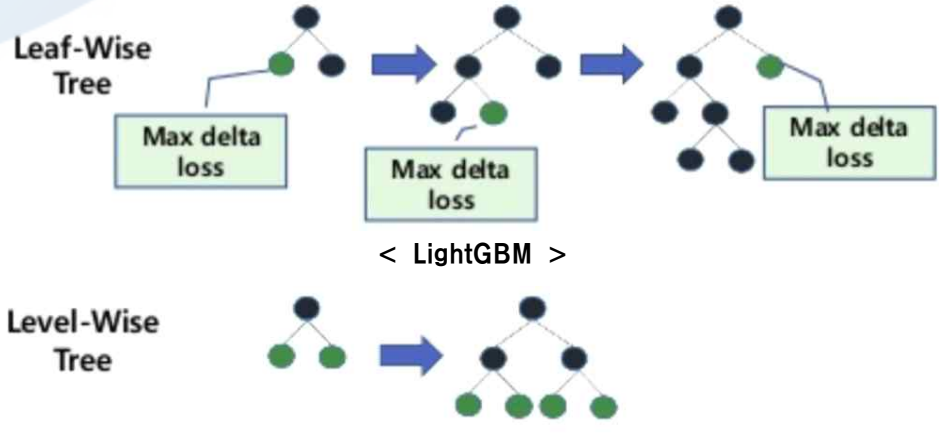


[그림 2-11] 2개로 분리한 잎(leaf)의 점수

2.2.3.3 LightGBM(Light Gradient Boosting Machine)

LightGBM은 Ke et al(2017)이 소개한 모델로, 의사결정트리(Decision Tree) 기반의 빠른 속도로 분류하는 고성능 그레디언트 부스팅(Gradient Boosting) 모델이다. 부스팅(Boosting)은 배깅(Bagging) 방법의 변형으로 모델이 잘 예측하지 못하는 부분을 개선하기 위한 방법으로, 여러 개의 모델을 순차적으로 학습시키며, 이전 모델이 잘 예측하지 못한 데이터들에 가중치를 부여하여 다음 모델에서 가중치가 추가된 데이터로 학습하는 방법이다(Ke et a.l, 2017).

[그림 2-12]에서 보는 것처럼 다른 부스팅 기반 모델(그림 아래)은 의사결정트리의 노드(node)들이 뿌리 노드(root node)와 가까운 노드를 우선적으로 순회하고 수평 성장하는 레벨 분할 방식(level wise, depth-first)을 사용한다. 반면 LightGBM(그림 위)은 최대 델타 손실(Max delta loss)이 큰 노드에서 분할하여 수직 성장하는 리프 분할 방식(leaf wise, best first)을 사용한다. 동일한 리프(leaf)를 성장시킬 때 리프 분할 방식(leaf wise)이 레벨 분할 방식(leaf wise)보다 더 많은 손실을 줄일 수 있다(Ke et al, 2017).



[그림 2-12] LightGBM과 다른 부스팅 모델의 트리 분할 방식

출처: https://www.slideshare.net/freepsw/boosting-bagging-vs-boosting

LightGBM은 높은 속도로 인해 'Light'으로 시작한다. LightGBM은 적은 메모리를 사용하여 대용량 데이터를 처리할 수 있으며, 레벨 분할 방식(leaf wise)보다 더 많은 손실을 줄일 수 있기 때문에 예측 정확도 또한 더 높다. LightGBM 은 병렬 GPU 학습도 지원하여 데이터 과학 응용 프로그램 개발에 LGBM을 널리 사용되고 있다. 하지만 LightGBM은 과적합(over-fitting)에 민감하여 소규모 데이터셋에는 사용하는 적절하지 않으며, 최소한 행(row)의 수가 10,000개 이상이 데이터에서 사용하는 것이 좋다[7].

LightGBM은 대용량의 데이터를 빠르고 정확하게 학습하기 위한 방법으로 GOSS(Gradient Based One Side Sampling)와 EFB(Exclusive Feature Bundling) 두 가지를 모두 사용한다. GOSS는 학습에 사용되는 데이터의 수를 줄이는 방법이며, EFB 예측변수의 수를 줄이는 방법이다. GOSS는 학습 데이터의 수를 줄이기 위해 기울기(gradient)를 사용한다. 경사부스팅에서 데이터의 기울기가 작으면 실제값과 예측값의 차인 오차가 작다는 해석이 가능하여 학습이 잘 된 것으로 판단한다. GOSS는 데이터의 기울기 절대값을 이용하여 학습 데이터를 정렬시킨 후, 절대값이 작은 데이터는 제거하여 학습에 사용되는 데이터의 개수를 줄인다. GOSS는 기울기가 작은 학습 데이터에 가중치를 부여하여 데이터 일부가 제외 되더라도 데이터 전체의 분포 변화를 최소화하는 방식으로 학습을 진행한다. LightGBM은 히스토그램 기반 알고리즘을 사용한다. 히스토그램 기반 알고리즘이란 연속형 독립변수의 데이터를 구간별로 나누고 구간별 정보량을 계산하여 각 트리의 노드를 분리하는 방법을 말한다. 이 방법을 사용하면 전체 데이터를 정렬하여 계산하는 방법에 비해 메모리 사용을 줄일 수 있을 뿐 아니라 학습 속도도 훨신 더 빨라진다. XGBoost도 히스토그램 알고리즘 옵션이 포함되어 있다. EFB(Exclusive Feature Bundling)는 독립변수를 더 적은 개수로 줄이는 방법이다. 독립변수가 많고 데이터가 산발적인 형태를 띄면 변수들이 동시에 0값을 취하는 경우는 드물다. 이러한 변수들은 서로 배타적이거나 배타적인 특성을 지니기 때문에 변수들이 가지고 있는 정보 손실을 최소화하면서 하나의 변수로 결합할

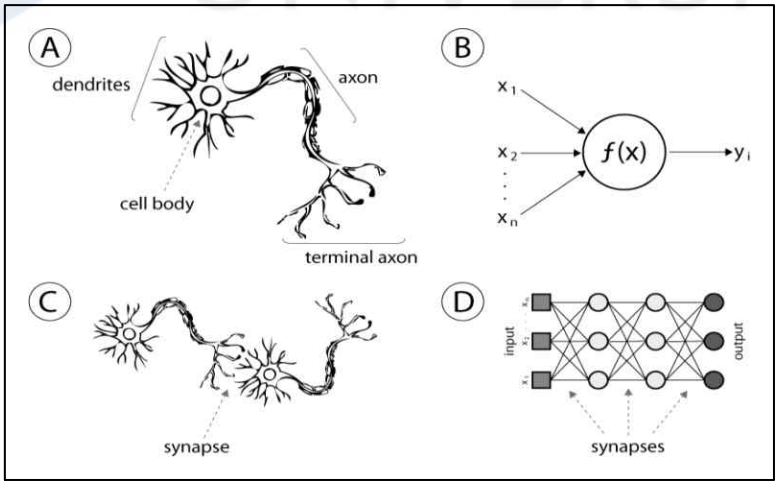
7) 출처: Pushkar Mandot(2017). https://medium.com/@pushkarmandot/https-medium-com-pushkarmandot-what-is-lightgbm-how-to-implement-it-how-to-fine-tune-the-parameters-60347819b7fc

수 있다. 여러 개의 독립변수를 하나의 하나의 변수로 결합하는 방법을 EFB(Exclusive Feature Bundle)라고 한다. 이렇게 하나로 결합한 독립변수에 히스토그램 기반의 알고리즘을 적용하여 학습을 진행한다. 예를 들면 변수 A의 구간이 [0, 10)이고, 변수 B의 구간이 [0, 20)이라면, 변수 A, B의 구간은 겹치지만 변수 B에 10을 더하면 [10, 30)이 되어 서로 다른 구간으로 분리가 가능하다. 이는 트리에 단조 변환(monotone transformation)을 주더라도 결과에는 영향을 미치지 않기 때문이다. 이제 두 변수의 묶음 값은 [0, 30)으로 동일한 구간을 갖도록 변형이 되었다. 이후 히스토그램 기반의 알고리즘을 적용하면 [0, 30)의 데이터 수보다 적은 구간으로 분할이 가능하며, 노드 분리 점을 계산하는 학습 속도도 빨라진다(김종영, 2019).

2.2.3.4 DNN(Deep Neural Networks, 심층신경망)

신경망(NN: Neural Network)은 인간의 두뇌 학습 과정을 뉴런과 노드의 망을 통한 연산 과정으로 간주하여 만든 모델이다. 인간이 경험으로 학습하는 두뇌의 신경망 활동, 즉 생물학적으로 뇌가 자극에 반응하는 원리를 활용하여 설계된 기법이다(이형탁, 2019). 신경망은 단순한 선형 패턴뿐 아니라 복잡한 비선형 패턴까지 학습할 수 있어 높은 예측성과를 보이고 있어 현재 데이터 분석 기법 중 가장 활발히 연구 및 활용되고 있다(박인근, 홍지후, 강남규, 김성호, 정구범, 2019).

신경망(NN: Neural Network)은 인간의 뇌에서 뉴런(Neuron)이 패턴을 인식하는 방식을 모방한 알고리즘이다. 뉴런은 신경망을 구성하는 각 신경세포를 지칭한다. 하나의 세포체(Cell body)에는 축색돌기(Axon)와 여러 개의 수상돌기(Dendrite)가 포함되어 있다. 신경세포들 간에 정보 전달은 시냅스(Synapse)라고 불리는 신경세포의 접합부를 통해 이루어진다. 인경신경망은 이러한 신경세포들이 모이면 인간의 지능과 유사한 형태의 학습이 가능하다는 원리에서 시작되었다(민성욱, 2017).

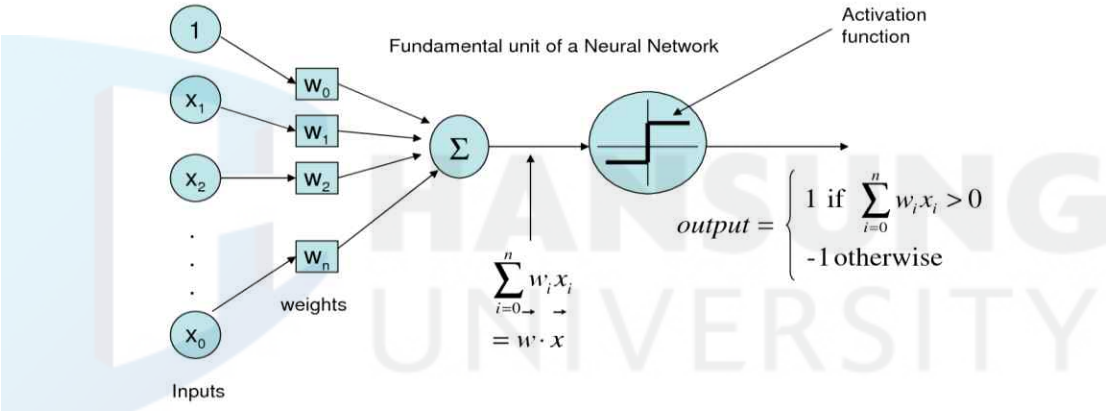


[그림 2-13] 인간의 뉴런 구조과 인공신경망 구조

출처: 민성욱(2017)

McCulloch과 Pitts(1943)는 생물학적 신경망 이론을 단순화하여 논리와 산술, 연산 기능 등을 구현할 수 있는 신경망 이론을 제시하였다. 이어 Hebb(1949)은 생물학적 신경망에서 신호가 발생할 때 나타나는 학습과 기억 효과를 인공신경망에도 적용할 수 있음을 보여주었으며, 이를 인공신경망에서 가중치(weight)라는 개념으로 설명하였다.

Rosenblatt(1958)은 Hebb의 가중치(weight) 개념을 적용한 [그림 2-14]와 같은 단층 신경망인 퍼셉트론(Perceptrons) 이론을 발표했다. 단층 퍼셉트론(SLP; Single Layer Perceptron)은 입력층(Input Layer)과 출력층(Output Layer)으로만 구성되어 있다.



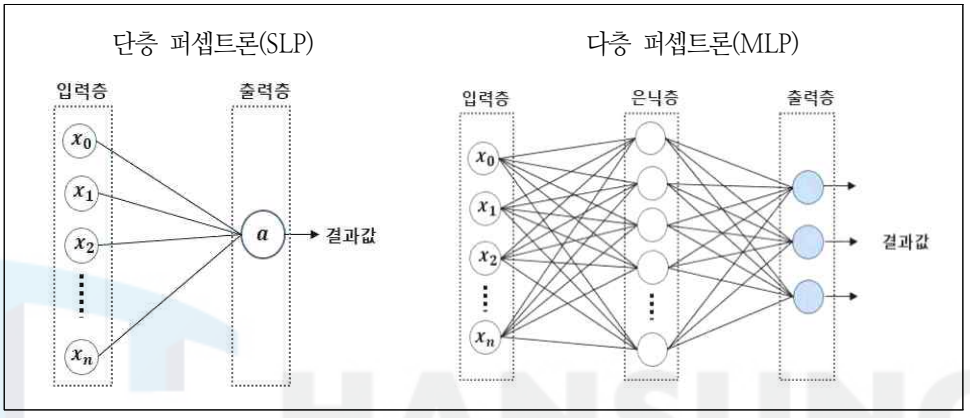
[그림 2-14] 퍼셉트론(Perceptron) 구조

출처: 민성욱(2017) 재인용

단층 퍼셉트론 구조에서 xj는 입력값으로 각 데이터의 특성(feature)을 나타낸다. wj는 각 입력값에 대한 가중치(weight)를 의미한다. 출력층은 입력값에 가중치를 곱한 값을 선형 결합한 뒤 지정된 활성화 함수(Activation function)를 통해 값의 부호를 조사하여 최종적으로 변수를 분류하거나 결과를 도출한다(원승현, 2017; 민성욱, 2017)

하지만 Minsky와 Papert(1969)는 Rosenblatt가 제시한 단층 퍼셉트론의 한계점을 발표하였다. 즉, 단층 퍼셉트론은 AND와 OR같은 선형적 분리 데이터에는 적용이 가능하지만, XOR 등 비선형적 분리 데이터에는 적용할 수 없다

는 것이다. 이에 [그림 2-12]와 같이 입력층과 출력층 사이에 하나 이상의 은닉층을 두어 비선형적 분리 데이터 학습이 가능한 다층 퍼셉트론(MLP; Multilayer Perceptron) 모델을 제시하였다(원승현, 2017). 하지만 이 발표를 계기로 신경망에 대한 불신과 함께 인공지능은 1970년대까지 침제기를 맞게 된다(박원기, 2018).



[그림 2-15] 단층 퍼셉트론(SLP)과 다층 퍼셉트론(MLP) 구조
출처: 원승현(2017)

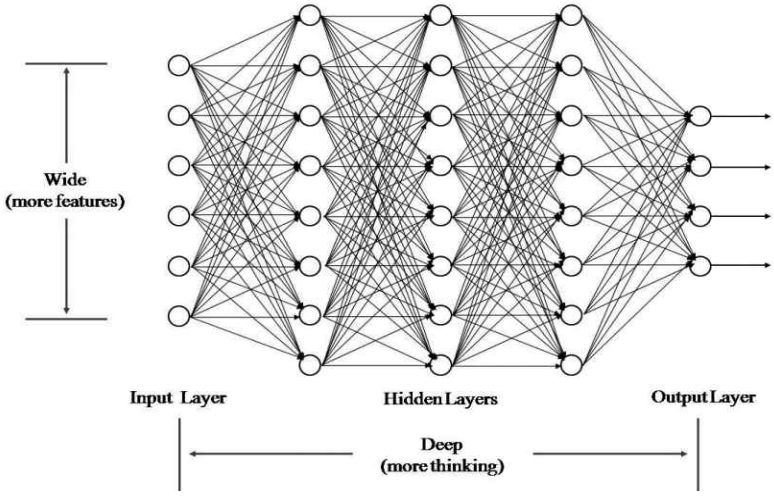
Werbos(1974)는 최초로 다층 신경망(MLP)의 학습 알고리즘으로 역전파 이론을 적용할 수 있음을 증명하였다. 하지만 당시의 인공지능에 대한 학계의 냉랭한 분위기로 인해 8년 후에서야 다층 신경망 내용을 학술지에 공개된다. 이 연구는 Parker와 LeCun(1985)가 재발견하면서 인공신경망은 새로운 국면을 맞이한다.

1980년대 역전파(Backpropagation) 알고리즘이 소개되면서 인공신경망 연구는 다시 주목을 받았고, 영상인식, 음성인식 등 다양한 분야에 적용되기 시작했다. 인공신경망은 일반적으로 은닉층과 노드 개수를 늘려 신경망 구조를 복잡하게 만들수록 성능이 좋다. 하지만 수없이 많이 연결되어 있는 뉴런 사이의 연결선에 파라미터(parameter)를 통해 적절한 가중치를 할당해야 하는 어려움이 있고, 알고리즘 처리 시간의 증가 및 과적합(over-fitting) 문제가 제기되면서 인공신경망은 다시금 사람들의 관심에서 벗어나게 되었다.

2006년 Hinton, Osindero, Teh가 네이처에 논문을 발표하면서 역전파 알고리즘이 학계에 본격적으로 알려졌다. Hinton 팀은 인공신경망의 각 층을 먼저 비지도 학습(Unsupervised Learning) 방식을 이용하여 전처리를 진행하고, 전처리한 자료를 여러 층으로 쌓아올리는 방식으로 인공신경망 최적화 문제를 해결하였다. 이를 계기로 정체 상태에 있었던 인공신경망 연구가 다시 활기를 띠게 되었다(Hinton, Osindero & Teh, 2006).

이후 2012년 힌튼 교수가 이끄는 팀이 이미지 인식 경연대회에서 우승하면서 심층신뢰망(DBN; Deep Belief Network) 이론의 효과를 검증하였다. DBN을 통해 DNN에서 역전파 알고리즘이 갖는 문제점을 대부분 해결 가능하게 되었다. 또한 기존 심층 신경망의 과적합 문제를 해결할 수 있는 방안으로 신경망의 각 층에서 노드의 일부만을 사용하는 드롭아웃(Dropout)기법이 고안되었으며, 기존에 활성함수로 사용하던 시그모이드 함수와 하이퍼볼릭 탄젠트 함수를 대체하여 ReLU(Rectified Linear Unit)를 사용함으로 심층신경망에서의 'Vanishing Gradient Problem'도 해결할 수 있게 되었다(민성욱, 2017)

DNN(Deep Neural Network, 심층신경망)은 [그림 2-16]과 같이 입력층(Input Layer)과 출력층(Output Layer) 사이에 복수 개의 은닉층(Hidden Layer)으로 이뤄진 인공신경망(ANN: Artificial Neural Network)을 말한다(Bengio, Courville & Vincent, 2013).



[그림 2-16] 심층신경망(DNN; Deep Neural Network)의 구조
출처: 심재헌(2016)

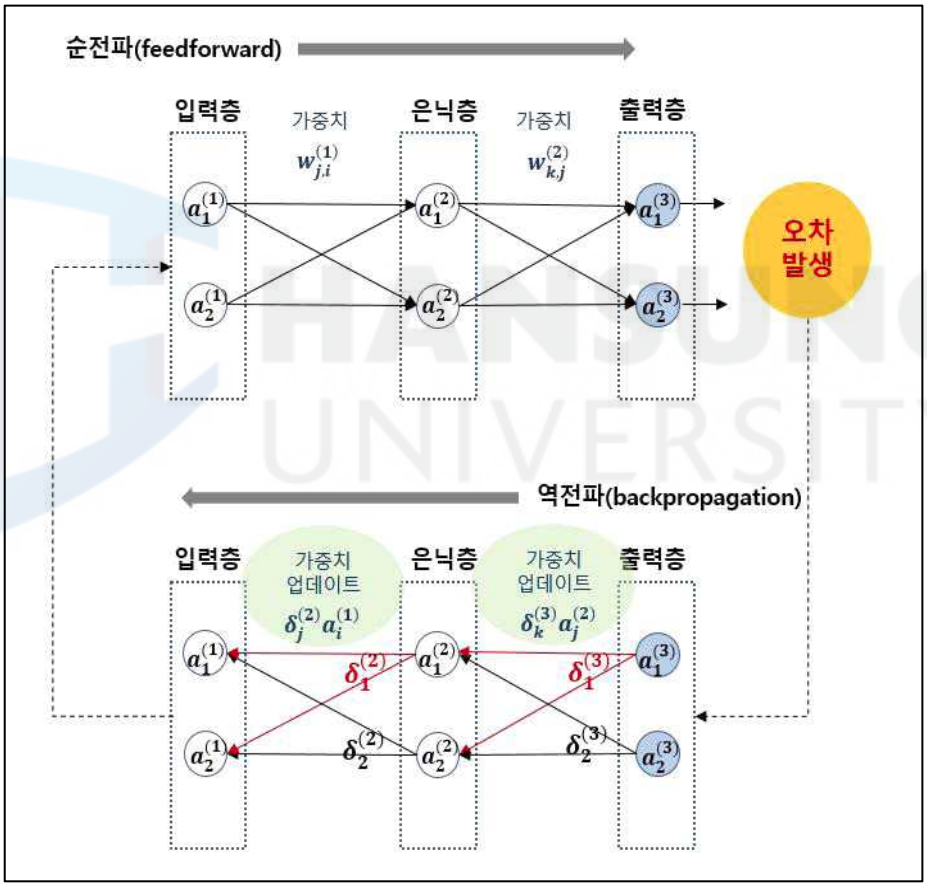
딥러닝(Deep Learning) 하면 주로 DNN(Deep Neural Network, 심층신경망)을 말한다. 딥러닝은 컴퓨터가 데이터 자체를 신경망 구조를 통해 학습하는 머신러닝의 특정한 한 분야라고 할 수 있다(이형탁, 2019) 딥러닝의 딥(deep)이라는 단어는 신경망을 구성하는 층(layer)의 개수가 많음을 의미한다. 2-3개의 층으로 구성된 인공신경망 구조를 쉘로우 러닝(Shallow Learning)이라고 하며, 층의 개수가 이보다 많으면 딥러닝이라 칭한다. 이와 같은 맥락에서 딥러닝을 심층신경망(DNN; Deep Neural Network이라고도 한다(심재헌, 2016)

현재 영상처리 패턴인식 등 다양한 분야에 사용되고 있는 딥러닝 모델에는 영상, 이미지와 같은 데이터를 처리하는 CNN(Convolutional Neural Network, 컨볼루션 신경망)과 음성인식 및 음악, 시퀀스가 있는 문자열 데이터를 처리하는 RNN(Recurrent Neural Network, 순환 신경망) 등이 대표적으로 사용되고 있으며, 각 분야별로 특화된 기법으로 진화하며 인공지능의 발전을 주도하고 있다(박원기, 2018)

다중 퍼셉트론에서 순전파(Feedforward)를 통해 도출된 오류를 역전파 알고리즘(Back Propagation)을 통해 해결하였다. 다층 퍼셉트론에서 정보의 흐름은 입력층에서 시작해 은닉층을 거쳐 출력층으로 진행되어 예측값을 도출하는데, 정보를 전방으로 전달하는 알고리즘을 순전파(Feedforward)라고 한다(원승현, 2017). 순전파에서 입력층의 노드 개수는 입력 데이터의 특성 개수와 일치해야 한다. 은닉층의 각 노드는 퍼셉트론 알고리즘에서와 같이 앞 층에서 전달 받은 데이터별로 해당 가중치를 선형적으로 곱하여 모두 합산하며, 합산된 값을 활성화 함수에 적용해 활성화 정도를 다음 층의 뉴런으로 전달하는 순전파를 수행한다(이용주, 2018). 특히 은닉층의 활성화 함수는 임계치를 적용해 활성화 정도를 부여하여 실질적으로 의미 없는 데이터는 사전에 필터링하는 역할을 수행한다(원승현, 2017). 출력층은 예측값의 유형에 따라 분류 또는 회귀를 수행한다. 반면 역전파(Back Propagation) 알고리즘은 신경망에서 가장 많이 사용되고 있는 학습 알고리즘으로 명확한 수학적 이론을 기반으로 하고 계산이 편리하다는 장점을 가지고 있다(원승현, 2017).

역전파 알고리즘은 예측값과 실제값의 차인 오차의 역전파를 통해 가중치

를 구한다. 역전파(Back Propagation)를 이용한 가중치 업데이트 절차는 다음과 같다. 1단계에서는 주어진 가중치를 적용하여 예측값을 계산한다(순전파). 2단계에서는 예측값과 실제값의 차인 오차를 구하며, 오차의 가중치는 경사하강법(Gradient Descent)으로 찾는다(오차의 역전파 계산). 3단계에서는 모든 가중치에 대해 2단계를 수행한다. 4단계에서는 1~3단계를 주어진 학습 횟수만큼 반복한다. 역전파(Back Propagation)를 이용한 가중치 업데이트 절차 도식화하면 [그림 2-17]과 같다.



[그림 2-17] 역전파 알고리즘 원리

출처: 원승현(2017)

역전파 알고리즘에서 오차를 최소화하는 가중치를 찾을 때 가장 널리 사용되는 최적화 알고리즘이 바로 경사하강법(Gradient Descent)이다(원승현, 2017). 경사하강법은 오차함수의 해당 가중치에 대한 기울기(경사)를 구하여 기울기가가 낮은 쪽으로 계속 이동하면서 최소값에 이를 때까지 반복하는 것이다. 이동거리와 오차함수의 θ_t에 대한 기울기(Gradient) 수식은 각각 (2-25), (2-26)과 같다.

$$v_{t+1} = \alpha f'(\theta_t) \quad \cdots (2-25)$$

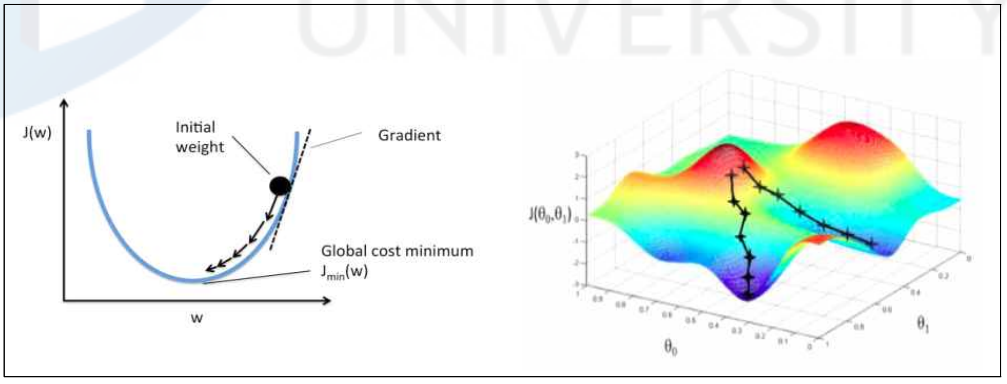
$$\theta_{t+1} = \theta_t - v_{t+1} \quad \cdots (2-26)$$

v_{t+1} : 이동거리

$f'(\theta_t)$: 오차함수 θ_t에 대한 기울기

θ_{t+1} : 오차함수 다음 기울기 측정 지점(가중치)

α : 학습률(Learning Rate) 또는 Step Size



[그림 2-18] 경사하강법(Gradient Descent) 개념도
출처: 원승현(2017)

경사하강법(Gradient Descent)은 네트워크의 파라미터(parameter)들을 θ라고 했을 때, 네트워크 추정치와 실제 결과값의 차인 오차함수 값을 최소화하기 위해 기울기를 이용하는 방법이다. 경사하강법에서는 θ에 대해 기울기

의 반대 방향으로 일정한 크기만큼 이동하는 것을 반복하여 오차함수 값을 최소화하는 θ의 값을 찾는다. 즉 기울기가 양수(+)이면 음수(−) 방향으로 이동하고, 음수(−)이면 양수(+) 방향으로 이동한다. 이때 α는 학습률(Learning Rate) 또는 step size로 불리며, 보통 0.01~0.001 정도를 사용한다. 경사하강법에는 Batch Gradient Descent, SGD(Stochastic Gradient Descent), Momentum, AdaGrad(Adaptive Gradient), RMSProp, ADAM(Adaptive Moment Estimation) 등이 있다(원승현, 2017).

활성화 함수(Ativation Function)는 은닉층의 노드에서 데이터와 가중치를 선형적으로 곱하여 모두 합산한 값에 임계치를 적용하여 활성화 정도를 부여함으로써 실질적으로 의미있는 데이터와 의미 없는 데이터를 필터링하는 역할을 수행하며, 비선형 데이터 처리 및 역전파 학습의 성능 향상을 위해 계속 발전되어 왔다(원승현, 2017). 활성화 함수에는 Sigmoid, tanh, ReLU(Rectified Linear Unit) 등이 있다. Sigmoid 함수는 선형 함수 결과를 0에서 1까지의 비선형 형태로 변환하는 함수이다. tanh 함수는 선형 함수의 결과를 −1에서 1까지의 비선형 형태로 변환하는 함수이다. ReLU(Rectified Linear Unit) 함수는 최근에 가장 많이 사용되는 활성화 함수로서, 0보다 높은 입력값은 그대로 출력하고 대신 0이하는 0을 출력한다(원승현, 2017).

신경망은 지금까지의 비정형 데이터라고 불리는 문자, 그림, 소리 데이터를 분석하는데 많이 활용되었다. 마케팅에서도 고객 댓글이나 SNS 상에서 자주 언급되는 단어와 내용에 대한 분석, 우수고객 확인 등에 많이 쓰인다. 신경망은 빅데이터를 분석을 위해 활용되고 있는 Data Analytics의 한 가지 방법이라고 생각할 수 있다. 마케팅에서의 빅데이터는 비정형 자료뿐 아니라 숫자로 표현되는 많은 자료들도 회사의 내부의 거래기록이나 소비자들과의 접촉기록, 또 관련된 경기나 날씨를 포함한 외부의 다양한 자료들이 해를 거듭할수록 쌓인다. 한글을 포함한 문자 자료나 비정형 자료들도 많지만, 숫자로 표현할 수 있는 빅데이터를 통해서 보다 정교한 정보를 만들어낼 필요가 있다(김주영, 2018)

2.3 선행연구와의 차별성

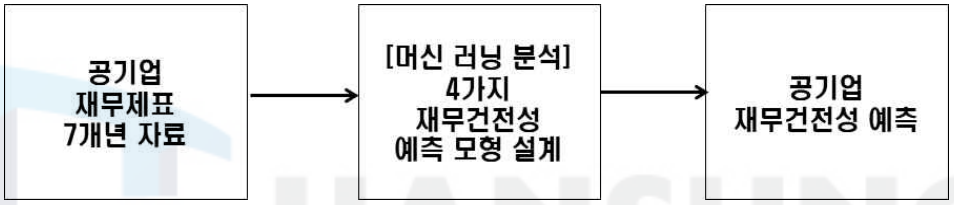
최근 몇 년 사이 머신러닝 기법의 급속한 발달과 함께 다양한 분야에서 머신러닝 기법을 활용한 연구가 진행되고 있다. 하지만 컨설팅 분야에서 머신러닝 기법을 이용한 연구는 찾아보지 못하였다. 이에 본 연구에서는 현재 기업에서도 머신러닝에 대한 관심이 증가하고 있는 시점에서 머신러닝 기법을 이용한 공기업 재무건전성 예측모델 실증연구를 통해 컨선턴트들에게 컨설팅 현장에서 머신러닝 기법을 적용할 수 있는 방안들을 모색해보고자 하였다.

또한 예측모델 설계 과정과 결과를 통해 도출된 요인의 분석을 통해 전통적 통계분석과 비교하여 머신러닝 기법의 특징과 함께 컨설턴트들이 머신러닝 기법을 실제 컨설팅 현장에서 적용시 고려해야 할 부분들을 제시하였다는 점에서 차별점이 있다.

III. 연구 설계

3.1 연구 모형

본 연구에서는 [그림 3-1]과 같이 공기업 7개년 재무제표 자료를 가지고 4가지 머신러닝 기법(Random Forest, XGboost, LightGBM, DNN)을 이용하여 공기업 재무건전성 예측모델을 만들어 도출된 예측값과 실제값을 비교해보고자 하였다.



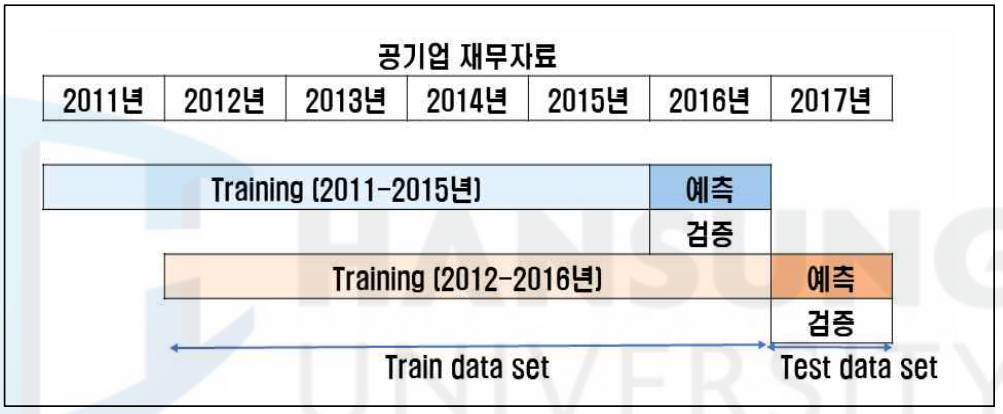
[그림 3-1] 연구 모형

3.2. 연구 자료

본 연구에서는 2011년부터 2017년까 7개년 연속 선정된 26개 공기업 제무재표 자료를 분석에 사용하였다. 2011년도는 한국채택국제회계기준(Korea-International Financial Reporting Standards, 이하 K-IFRS) 도입 연도로 공공기관의 회계기준이 바뀐 첫 해이다. 이에 K-IFRS가 적용되는 첫 해인 2011년부터 2017년까지의 공기업 재무자료를 가지고 공기업 재무건전성 예측모델을 설계하였다. 해당 공기업 재무제표 자료는 기획재정부 출자관리과에 정보공개 요청을 통해 얻었으며, 일부 재무제표 자료는 해당 기업 홈페이지를 통해 수집하였다.

3..3 연구 방법

공기업 재무건전성 예측모델의 학습 및 예측 목표 기간은 [그림 3-2]와 같다. 5개년 단위로 모델을 학습시킨 후, 다음 6번째 연도의 재무건전성 값을 예측하도록 하였다. 즉, 2011-2015년, 2012-2016년 공기업 재무 데이터를 가지고 학습 후 각각 2016년과 2017년 공기업 재무건전성 예측값을 도출하여 실제값을 비교하였다. 또한 모델 설계 결과 도출된 요인들의 특징을 분석하여 제시하였다.



[그림 3-2] 공기업 재무건전성 예측모델링 형태

3.4 모델 평가

회귀 예측모델에서 예측에 대한 성능 평가는 주로 오차함수(Error Function) 값으로 확인한다. 본 연구에서는 MAE[8](Mean Absolute Error, 평균절대오차), MSE[9](Mean Squared Error, 평균제곱오차), RMSE[10](Root Mean Squared Error, 평균제곱근오차), MAPE[11](Mean Absolute Percentage Error, 평균절대비율오차) 4개의 오차함수 값으로 모델의 성능을 확인하였다. MAPE의 경우, 다른 평가지표와 달리 백분율(%)로 표기한다. 예측모델의 오차함수 값이 작을수록 예측 성능이 좋은 것으로 평가한다.

예측모델에서 주로 가장 많이 쓰이는 성능 평가 지표는 MSE와 RMSE이며, 수치 예측모델에서는 MAPE도 자주 사용된다. 이에 본 연구에서 결과를 제시할 때는 4가지 평가 지표값을 다 제시하되, 해석 시에는 주로 MSE나 RMSE, MAPE를 가지고 예측모델의 성능에 대해 정리하였다.

MAE(Mean Absolute Error, 평균절대오차)는 모든 절대 오차의 평균을 의미하며, 수식은 (2-27)과 같다.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\hat{y_i} - y_i\right| \quad \cdots (2\text{-}27)$$

MSE(Mean Squared Error, 평균제곱오차)는 부호의 영향을 제거하기 위해 절대값이 아닌 제곱을 사용하며, 수식은 (2-28)과 같다. MAE(Mean Absolute Error, 평균절대오차)와는 절대값을 쓰는지, 제곱을 쓰는지의 차이다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y_i} - y_i)^2 \quad \cdots (2\text{-}28)$$

8) MAE : 위키피디아, https://en.wikipedia.org/wiki/Mean_absolute_error
9) MSE: 위키피디아, https://en.wikipedia.org/wiki/Mean_squared_error
10) RMSE: 위키피디아, https://en.wikipedia.org/wiki/Root-mean-square_deviation
11) MAPE: 위키피디아, https://en.wikipedia.org/wiki/Mean_absolute_scaled_error

수식 (2-27), (2-28)에서 n은 데이터 수, $\hat{y_i}$는 예측값, y_i는 실제값을 의미한다.

RMSE(Root Mean Squared Error, 평균제곱근오차)는 MSE(Mean Squared Error, 평균절대오차)에 제곱급을 취한 값이며, 수식은 (2-29)와 같다. n은 데이터 수, $\hat{y_i}$는 예측값, y_i는 실제값을 의미한다.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y_i}-y_i)^2} \quad \cdots (2\text{-}29)$$

MAPE(Mean Absolute Percentage Error, 평균절대비율오차)의 예측 방법의 예측 정확도를 측정한 것으로 머신러닝의 회귀문제에 대한 손실함수로 사용된다. MAPE는 주로 백분율로 정확도를 표시하며, 수식은 (2-30)과 같다. 여기서 n은 데이터 수, $\hat{y_i}$는 예측값, y_i는 실제값을 의미한다.

$$MAPE = 100\% \times \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{y_i}-y_i|}{|y_i|} \quad \cdots (2\text{-}30)$$

MAPE는 일반적으로 상대 오차 측면에서 매우 직관적인 해석으로 인하여 회귀 문제 및 모델 평가에서 손실 함수로 사용된다. MPAE는 수치를 예측하는 모델에서 주로 사용된다. RMSE가 눈금 의존적 오차라면 MAPE는 비율 의존적 오차로, 수치형 데이터의 경우 MAPE 값이 더 정밀도가 있다고 본다.

IV. 실증 분석

4.1 변수 선정

본 연구에서는 공기업 재무건정성 예측모델 설계를 위해 목표변수를 재무건전성 1개 변수로 설정하고, 독립변수는 재무제표 관련 데이터로 총 27개의 변수를 설정하였다.

4.1.1 목표변수 설정

목표변수인 재무건전성은 총 11개 지표를 가지고 산출한 재무건전성지수를 인 KJY Score[12]로 설정하였다. Piotroski가 제시한 9가지 변수에 고덕필(2003), 박경덕, 한길석, 윤석진(2008)의 연구를 종합하여 이자보상배율과 자산규모의변화 2가지 변수를 추가하여 〈표 4-1〉과 같이 11개 지표로 구성된 재무건전성지수인 KJY-Score 산출식을 만들었다.

〈표 4-1〉 KJY Score 산출식

산출식	$KJY\ Score = K_1 + K_2 + K_3 + K_4 + K_5 + K_6 + K_7 + K_8 + K_{10} + K_{11}$	
산출 지표	K_1: 총자산경상이익률	K_7: 유상증자 여부
	K_2: 총자산경상이익률의 변화	K_8: 매출액 영업이익률의 변화
	K_3: 총자산대비영업현금흐름비율	K_9: 자산회전율의 변화
	K_4: 영업현금흐름과 경상이익의 차이	K_{10}: 이자보상배율
	K_5: 레버리지 변화	K_{11}: 자산규모의 변화
	K_6: 유동비율의 변화	

출처: 정준수(2017) 산출식 인용, 일부 수정

12) KJY Score: 11개 재무지표를 통해 산출된 재무건전성지수를 의미한다. 정준수(2017)가 2011회계연도 1개년 공공기관 재무자료를 가지고 11개 지표로 산출된 재무건전성지수를 도출하여 단순화된 지수로 공공기관 재무적 성과 측정 지표로서의 활용 가능성을 검증하였다. 이어 본 연구자(윤혜란)가 2011부터 2017년까지 7개년 회계연도 공기업 재무 데이터를 가지고 11개 지표로 산출된 재무건전성지수를 도출하여 공기업 재무적 성과에 관한 대표성을 재검증하였다. 한편 이 11개 지표로 산출에 대해 김상봉 교수가 처음 제안하고 자문을 제공하여 이 11개 지표로 산출된 재무건전성지수를 세 사람의 영문 성을 각각 따서 KJY Score라고 명명하였다.

〈표 4-2〉 KJY Score 구성 11가지 지표와 계산식

No	KJY Score 지표	계산식	점수 부여	
			1점	0점
1	총자산경상이익률	경상이익/기초자산	0초과	0이하
2	총자산경상이익률의 변화	당기총자산경상이익률-전기총자산경상이익률	0초과	0이하
3	총자산대비 영업현금흐름비율	영업활동으로 인한 현금흐름/기초자산총액	0초과	0이하
4	영업현금흐름과 경상이익의 차이	(영업활동으로 인한 현금흐름-경상이익) /기초자산총액	0초과	0이하
5	레버리지 변화	(기말장기금융부채-기초장기금융부채) /평균자산총액 레버리지=금융부채/총자산	0미만	0이하
6	유동비율의 변화	당기유동비율-전기유동비율 (유동비율=유동자산/유동부채)	0초과	0이하
7	유상증자 여부	유상증자의 여부	유상증자 없음	유상증자 있음
8	매출액 영업이익률의 변화	당기매출액영업이익율-전기매출액영업이익율 (매출액영업이익율=영업이익/매출액	0초과	0이하
9	자산회전율의 변화	당기자산회전율-전기자산회전율 (자산회전율=매출액/기초자산)	0초과	0이하
10	이자보상배율	영업이익/이자비용	1이상	1미만
11	자산규모의 변화	당기자산규모-전기자산규모 (자산규모=Ln(총자산)) *ln: 자연로그	0초과	0이하

출처 : 정준수(2017) 표 인용, 일부 수정

KJY Score는 〈표 4-2〉에서 처럼 각 지표에 해당하는 점수 부여 기준에 따라 1점 또는 0점을 부여하여 총 합산 점수로 산출한다. KJY Score는 총 11개의 지표를 사용하여 이를 종합하여 반영한 지표로, KJY Score의 최소값은 0, 최대값은 11점이다.

한편 공기업 KJY Score 산출과정 세부항목 계산식은 〈표 4-3〉과 같다.

〈표 4-3〉 공기업 KJY Score 산출과정 세부항목 계산식

세부항목	계산식
장기금융부채	장기차입금 + 사채
경상이익	영업이익손실 + 기타이익손실
기초자산	(자산총계 + 전기자산총계)의 평균
총자산경상이익률	경상이익 / 기초자산
총자산경상이익률의 변화	총자산경상이익률 - 전기총자산경상이익률
총자산 대비 영업현금흐름비율	영업활동으로인한현금흐름 / 기초자산
영업현금흐름과 경상이익의 차이	(영업활동으로인한현금흐름 - 경상이익) / 기초자산
레버리지변화	(장기금융부채 - 전기장기금융부채) / 기초자산
유동비율	유동자산 / 유동부채
유동비율의 변화	유동비율 - 전기유동비율
매출액영업이익률	영업이익손실 / 수익매출액 *영업이익손실 변수가 없고 영업이익 변수가 있는 경우: 영업이익 / 수익매출액 *수익매출액 변수가 없고 영업수익 변수가 있는 경우: 영업이익손실 / 영업수익
매출액영업이익률의 변화	매출액영업이익률 - 전기매출액영업이익률
자산회전율	수익매출액 / 기초자산 *수익매출액 변수가 없고 영업수익 변수가 있는 경우: 영업수익 / 기초자산
자산회전율의 변화	자산회전율 - 전기자산회전율
이자보상배율	영업이익손실 / 이자비용[13] *영업이익손실 변수가 없고 영업이익 변수가 있는 경우: 영업이익 / 이자비용
자산규모의 변화	ln(자산총계) - ln(전기자산총계) ※ ln : 자연로그

출처 : 정준수(2017) 표 인용, 일부 수정

〈표 4-3〉의 공기업 KJY Score산출과정 세부항목 계산식, 〈표 4-2〉 KJY Score 지표와 계산식, 〈표 4-1〉의 KJY Score산출식을 통해 도출된 26개 공기업의 2011년도부터 2017년도까지의 7개년 KJY Score는 〈표 4-4〉와 같다.

13) 이자비용이 0이면 이자보상배율을 구할 수 없기 때문에 이자비용이 0인 경우에는 0.0001로 변환하여 계산한다(정준수, 2017).

〈표 4-4〉 7개년 회계연도 공기업 KJY Score (2011-2017)

NO	2011년	2012년	2013년	2014년	2015년	2016년	2017년
1	7	8	8	8	6	6	7
2	4	5	9	10	5	8	6
3	9	10	6	9	9	8	8
4	10	7	8	9	8	7	7
5	8	7	7	8	8	6	7
6	6	9	9	8	9	6	8
7	8	4	8	6	8	9	9
8	6	9	9	7	6	9	6
9	6	7	8	6	7	5	6
10	5	2	5	4	5	5	7
11	4	3	5	5	4	3	4
12	7	6	9	6	6	4	5
13	7	10	6	7	8	7	7
14	7	7	6	6	7	6	6
15	6	6	10	8	6	6	8
16	7	6	5	4	8	6	4
17	8	8	8	9	8	9	8
18	7	8	9	7	6	10	6
19	5	5	5	8	6	9	6
20	6	6	7	10	7	9	7
21	7	6	7	9	9	8	5
22	5	8	7	9	7	8	7
23	7	8	6	9	7	9	8
24	7	9	7	7	9	9	6
25	8	10	7	7	9	9	7
26	6	9	6	8	7	9	7

주) 본 연구에서는 각 기업별 특성을 파악하는 게 아니므로, 기업명 대신 번호를 부여함.

이렇게 2011년부터 2017년까지 7개년 회계연도 공기업 재무 데이터를 통해 도출된 KJY Score는 2016년과 2017년 공기업 재무건전성 예측모델에 설계하는 데 사용하였다. 7개년 KJY Score 중 2016년과 2017년 KJY Score는 2016년과 2017년 공기업 재무건전성 예측모델에서 도출된 예측값과의 비교검증에 사용하였다.

4.1.2 독립변수 설정

독립변수는 공기업 결산서 내 재무제표 계정과목과 계정과목을 가지고 일부 파생변수를 만들어 〈표 4-5〉와 같이 총 27개의 변수를 설정하였다. 파생변수의 산출값은 〈표 4-2〉와 〈표 4-3〉에 제시된 항목 계산식을 통해 도출하였다.

〈표 4-5〉 독립변수 설정 및 변수 표기

NO	독립변수	변수 표기	실제 변수표기
1	경상이익	Ordinary Income	OrdiI
2	기초자산	Underlying Assets	UA
3	기타이익손실	Other Income	OtherI
4	리버리지변화	Diff. of Leverage	DiffLev
5	매출액영업이익률	Operating Income/Revenue	OI/Rev
6	매출영업이익률변화	Diff. of Operating Income/Revenue	DiffOI/Rev
7	사채	Debenture	Debt
8	수익매출액	Revenue	Revenue
9	영업이익손실	Operating Income	OperI
10	영업현금흐름과 경상이익의차이	Operating Cash Flow - Ordinary Income	OCF-OrdiI
11	영업활동현금흐름	Operating Cash Flow	OCF
12	유동부채	Current Liabilities	CL
13	유동비율	Current Ratio	CR
14	유동비율변화	Diff. of Current Ratio	DiffCR
15	유동자산	Current Assets	CA
16	유상증자여부	Rights Issue	RI
17	이자보상배율	Interest Coverage Ratio	ICR
18	이자비용	Interest Cost	IC
19	자산규모의변화	Diff of Asset Size	DiffAsize
20	자산총계	Total Assets	TA
21	자산회전율	Assets Turnover	AT
22	자산회전율변화	Diff of Asset Turnover	DiffAT
23	장기금융부채	Long-Term Financial liabilities	LTFL
24	장기차입금	Long-Term Borrowings	LTB
25	총자산경상이익률	Ordinary Income/ Underlying Asset	OrdiI/UA
26	총자산경상이익률의변화	Diff. of Underlying Asset/ Underlying Asset	DiffOrdiI/UA
27	총자산대비영업현금흐름비율	Operating Cash Flow/ Underlying Asset	OCF/UA

4.2 데이터 전처리(Data Preprocessing)

머신러닝 분석 작업과 관련하여 데이터 전처리(Data Preprocessing)는 반드시 거쳐야 하는 과정이다. 분석 결과 및 인사이트 제공과 함께 모델 성능에 직접적인 영향을 미치기 때문이다(송현화, 2019).

데이터 전처리에는 데이터 정제, 데이터 통합, 데이터 변환, 데이터 축소 등이 있다. 데이터 정제는 결측치 및 이상값, 오류나 분산 등의 잡음이 섞인 데이터를 처리하는 것이다. 데이터 통합은 데이터 분석이 용이하도록 유사 데이터 및 연계가 필요한 데이터 또는 데이터 베이스(DB)들을 통합하는 기술을 말한다. 데이터 변환은 데이터 유형 변환 등 데이터 분석이 용이한 형태로 변환하는 기술로 정규화, 집합화, 요약, 계층 생성 등의 방법을 활용할 수 있다. 데이터 축소는 분석에 불필요한 데이터를 축소하여 고유한 특성은 손상되지 않도록 하고 분석에 대한 효율성을 증대하기 위한 것으로 차원축소나 데이터 압축, 주성분분석(PCA) 등이 있다[14].

본 연구에서는 데이터 변환 내용은 독립변수 설정에서 다루었으며, 본 절에서는 결측치(Missing value)와 변수 간 상관관계(Correlation)를 살펴보았다.

4.2.1 결측치(Missing Value) 확인

결측치를 처리하는 방식은 결측치를 제거하거나 수치형 데이터인 경우 경우 평균이나 중앙값으로, 범주형의 경우 최빈값으로 대체한다. 가장 쉬운 방법은 Null 포함 행 혹은 일부 행을 제거하는 것이며, 만약 샘플수가 충분하지 않은 경우 Null 값을 특정 값으로 채울 수 있다(송현화, 2019)

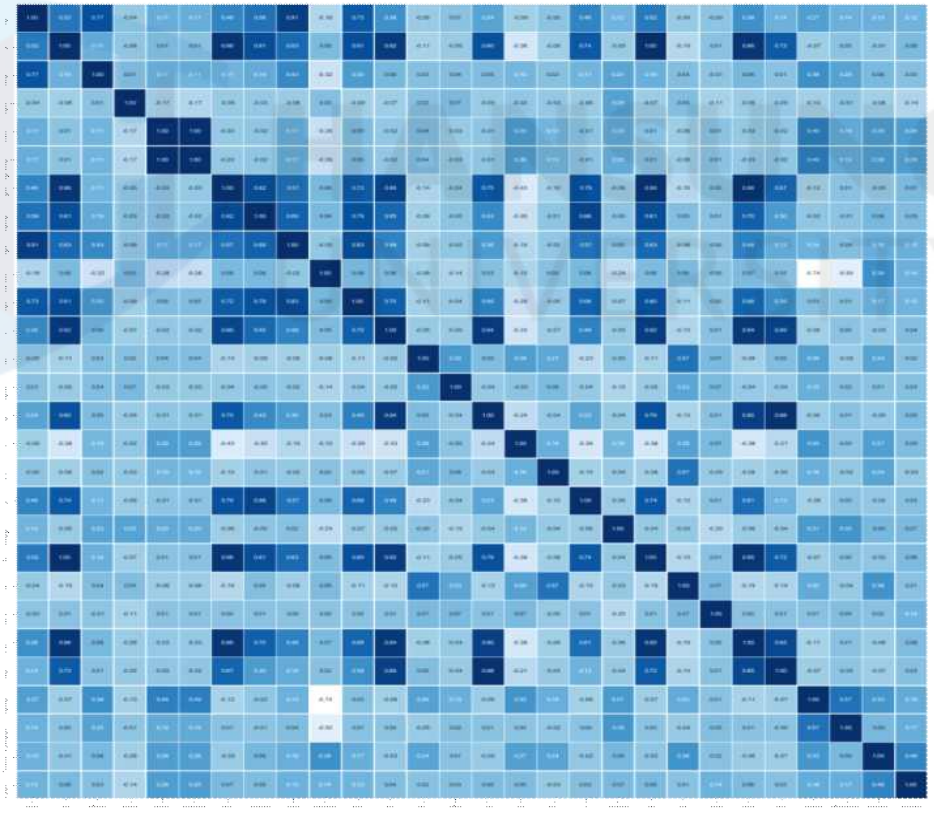
본 연구에서 결측치 확인 결과, 한 개의 열에서 4개 변수 해당 데이터에서 NaN(Not a Number)이 발견되어 Nan을 0으로 채워 결측치를 처리하였다.

14) 출처: https://www.saedsayad.com/data_preparation.htm

4.2.2 상관 분석(Correlation Analysis)

변수 간 관계 정도를 파악하기 위하여 상관관계(correlation)를 살펴보았다. 다음 페이지의 〈표 4-6〉은 상관분석 결과를 나타낸 것이다.

[그림 4-1]은 28개 변수 간 상관분석를 결과를 히트맵(heat map)으로 표현한 것이다. 히트맵(heat map)은 열을 뜻하는 히트(heat)와 지도를 뜻하는 맵(map)을 결합시킨 단어로, 색상으로 표현할 수 있는 다양한 정보를 일정한 이미지 위에 열 분포 형태의 비쥬얼한 형태의 그래픽으로 출력하는 것이 특징이다[15]. 히트맵(heat tmap)으로 표현된 상관관계 결과에서 색이 진할수록 높은 상관관계가 있음을 의미한다. 즉, 변수 간 유의성이 있음을 파악하였다.



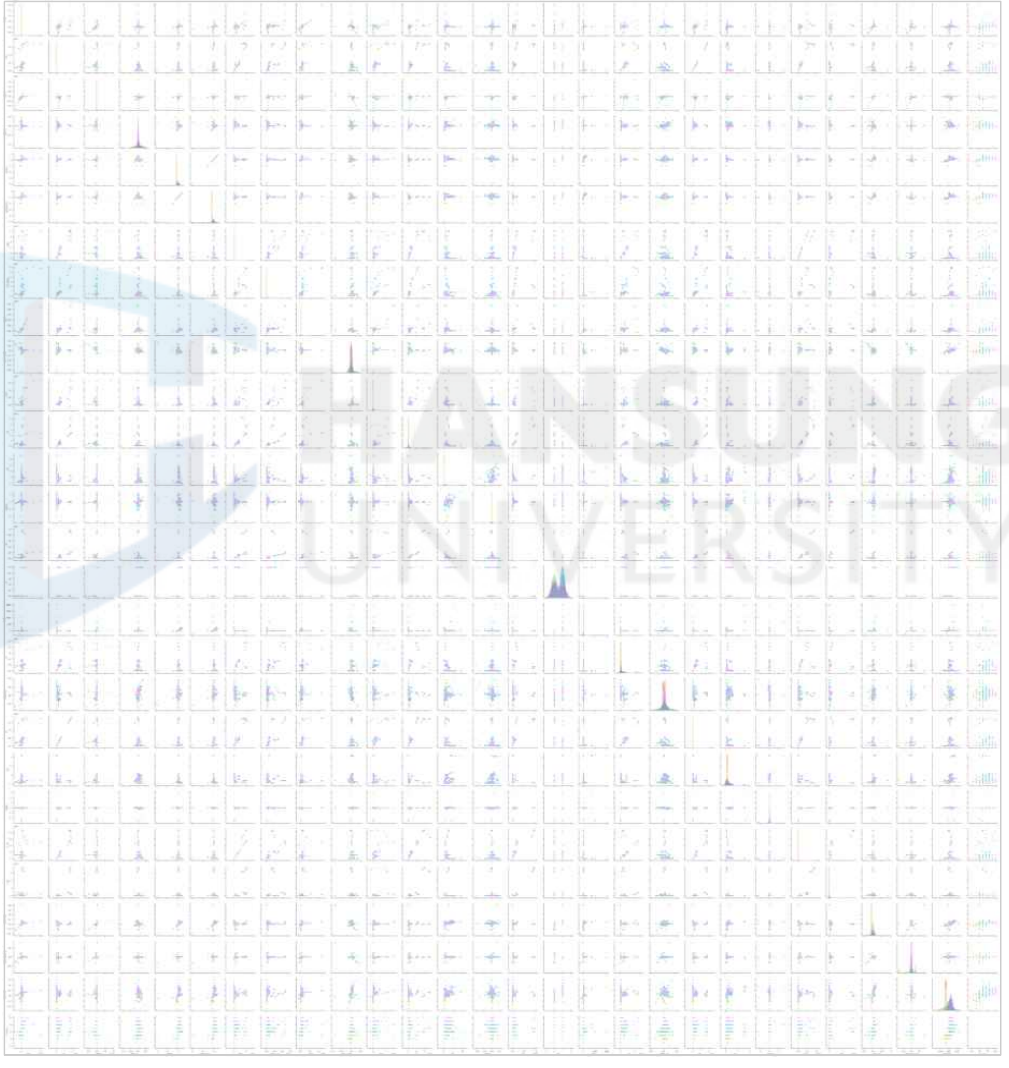
[그림 4-1] 히트맵(heat map)으로 나타낸 상관 관계

15) 출처 : 위키백과. https://ko.wikipedia.org/wiki/히트맵

〈표 4-6〉 상관분석 결과

	OrdiI	UA	OtherI	DiffLev	OI/Rev	DiffOI/Rev	Debt	Revenue	OperI	OCF-OrdiI	OCF	CL	CR	DiffCR	CA	RI	ICR	IC	DiffAsize	TA	AT	DiffAT	LTFL	LTB	OrdiI/UA	DiffOrdiI/UA	OCF/UA	F-score
OrdiI	1	0.516721	0.769156	-0.03985	0.169248	0.169248	0.460291	0.576169	0.908882	-0.16217	0.730011	0.377731	-0.05288	0.008202	0.238668	-0.08524	-0.00431	0.457717	0.120981	0.521191	-0.03774	-0.0006	0.376899	0.136127	0.273713	0.144057	0.148939	0.118605
UA	0.516721	1	0.147784	-0.07761	0.00576	0.00576	0.963699	0.810492	0.632564	0.06205	0.805613	0.918971	-0.11165	-0.05386	0.795726	-0.37985	-0.07998	0.738515	-0.05129	0.999773	-0.19023	0.009589	0.955236	0.726786	-0.07419	0.001431	-0.01491	0.06106
OtherI	0.769156	0.147784	1	0.006157	0.111412	0.111412	0.114643	0.19361	0.43255	-0.31738	0.299607	0.087927	0.028411	0.044672	0.051998	0.099814	0.023549	0.112586	0.233242	0.155649	0.043414	-0.00675	0.083887	0.007351	0.382212	0.248193	0.080075	0.033571
DiffLev	-0.03985	-0.07761	0.006157	1	-0.16911	-0.16911	-0.05419	-0.03315	-0.06025	0.034683	-0.08792	-0.07313	0.022367	0.074429	-0.05455	-0.02277	-0.03441	-0.06159	0.250715	-0.07466	0.042326	-0.11493	-0.0563	-0.04775	-0.0951	-0.01367	-0.08223	-0.14437
OI/Rev	0.169248	0.00576	0.111412	-0.16911	1	1	-0.03231	-0.01645	0.166073	-0.27785	0.053543	-0.02349	0.041942	-0.03451	-0.01118	0.287146	0.102877	-0.0065	0.19628	0.007118	-0.06168	0.00541	-0.03024	-0.01961	0.489619	0.19113	0.283646	0.288461
DiffOI/Rev	0.169248	0.00576	0.111412	-0.16911	1	1	-0.03231	-0.01645	0.166073	-0.27785	0.053543	-0.02349	0.041942	-0.03451	-0.01118	0.287146	0.102877	-0.0065	0.19628	0.007118	-0.06168	0.00541	-0.03024	-0.01961	0.489619	0.19113	0.283646	0.288461
Debt	0.460291	0.963699	0.114643	-0.05419	-0.03231	-0.03231	1	0.824401	0.574578	0.079669	0.723748	0.858797	-0.14156	-0.04081	0.747828	-0.43166	-0.10038	0.790936	-0.0624	0.964293	-0.19085	0.00075	0.960483	0.672432	-0.11788	0.009519	-0.05032	0.069147
Revenue	0.576169	0.810492	0.19361	-0.03315	-0.01645	-0.01645	0.824401	1	0.686529	0.077279	0.789223	0.645356	-0.08943	-0.00331	0.428056	-0.29942	-0.00738	0.876115	-0.00068	0.813516	0.046039	0.010311	0.697931	0.304674	-0.01971	-0.006	0.081563	0.051804
OperI	0.908882	0.632564	0.43255	-0.06025	0.166073	0.166073	0.574578	0.686529	1	-0.02168	0.834402	0.475535	-0.09314	-0.01758	0.302787	-0.1854	-0.02145	0.572289	0.01847	0.633738	-0.08158	0.003558	0.476997	0.187256	0.136731	0.04127	0.157872	0.145424
OCF-OrdiI	-0.16217	0.06205	-0.31738	0.034683	-0.27785	-0.27785	0.079669	0.077279	-0.02168	1	0.091533	0.05754	-0.08257	-0.14218	0.028786	-0.14665	0.016745	0.075397	-0.24094	0.057917	0.062435	0.002289	0.065388	0.023969	-0.74223	-0.50107	0.384247	0.143866
OCF	0.730011	0.805613	0.299607	-0.08792	0.053543	0.053543	0.723748	0.789223	0.834402	0.091533	1	0.703299	-0.10855	-0.04263	0.487551	-0.29495	-0.04577	0.675719	-0.06568	0.801742	-0.10575	0.001817	0.659749	0.392539	0.026539	0.013599	0.165329	0.119997
CL	0.377731	0.918971	0.087927	-0.07313	-0.02349	-0.02349	0.858797	0.645356	0.475535	0.05754	0.703299	1	-0.04879	-0.05383	0.936394	-0.33423	-0.06677	0.479487	-0.05295	0.916321	-0.15399	0.011138	0.94407	0.894487	-0.08317	0.004807	-0.03363	0.042032
CR	-0.05288	-0.11165	0.028411	0.022367	0.041942	0.041942	-0.14156	-0.08943	-0.09314	-0.08257	-0.10855	-0.04879	1	0.315524	0.049616	0.281695	0.213455	-0.22858	-0.00419	-0.11165	0.565171	0.070809	-0.08245	0.047132	0.26002	-0.05054	0.242055	0.016809
DiffCR	0.008202	-0.05386	0.044672	0.074429	-0.03451	-0.03451	-0.04081	-0.00331	-0.01758	-0.14218	-0.04263	-0.05383	0.315524	1	-0.04239	-0.00339	0.062999	-0.03535	-0.15388	-0.05499	0.229346	0.069189	-0.04455	-0.04168	0.153129	0.020323	0.01094	0.033279
CA	0.238668	0.795726	0.051998	-0.05455	-0.01118	-0.01118	0.747828	0.428056	0.302787	0.028786	0.487551	0.936394	0.049616	-0.04239	1	-0.2391	-0.04464	0.229493	-0.04069	0.792857	-0.12945	0.011292	0.899295	0.984228	-0.06224	0.007765	-0.04525	0.047654
RI	-0.08524	-0.37985	0.099814	-0.02277	0.287146	0.287146	-0.43166	-0.29942	-0.1854	-0.14665	-0.29495	-0.33423	0.281695	-0.00339	-0.2391	1	0.160636	-0.3841	0.101804	-0.37694	0.260614	0.068002	-0.3837	-0.20808	0.349091	0.000967	0.274621	0.049674
ICR	-0.00431	-0.07998	0.023549	-0.03441	0.102877	0.102877	-0.10038	-0.00738	-0.02145	0.016745	-0.04577	-0.06677	0.213455	0.062999	-0.04464	0.160636	1	-0.09654	-0.03743	-0.07999	0.567009	-0.04721	-0.08657	-0.04131	0.160338	-0.02497	0.244446	-0.03478
IC	0.457717	0.738515	0.112586	-0.06159	-0.0065	-0.0065	0.790936	0.876115	0.572289	0.075397	0.675719	0.479487	-0.22858	-0.03535	0.229493	-0.3841	-0.09654	1	-0.05919	0.74207	-0.14884	0.011332	0.60938	0.132171	-0.09305	0.002954	-0.02213	0.032824
DiffAsize	0.120981	-0.05129	0.233242	0.250715	0.19628	0.19628	-0.0624	-0.00068	0.01847	-0.24094	-0.06568	-0.05295	-0.00419	-0.15388	-0.04069	0.101804	-0.03743	-0.05919	1	-0.03972	-0.02901	-0.20425	-0.06054	-0.04358	0.308123	0.346136	0.085537	0.070859
TA	0.521191	0.999773	0.155649	-0.07466	0.007118	0.007118	0.964293	0.813516	0.633738	0.057917	0.801742	0.916321	-0.11165	-0.05499	0.792857	-0.37694	-0.07999	0.74207	-0.03972	1	-0.18927	0.008788	0.954371	0.723368	-0.07078	0.00356	-0.01603	0.061331
AT	-0.03774	-0.19023	0.043414	0.042326	-0.06168	-0.06168	-0.19085	0.046039	-0.08158	0.062435	-0.10575	-0.15399	0.565171	0.229346	-0.12945	0.260614	0.567009	-0.14884	-0.02901	-0.18927	1	0.069729	-0.18791	-0.14056	0.196803	-0.04346	0.35896	0.006886
DiffAT	-0.0006	0.009589	-0.00675	-0.11493	0.00541	0.00541	0.00075	0.010311	0.003558	0.002289	0.001817	0.011138	0.070809	0.069189	0.011292	0.068002	-0.04721	0.011332	-0.20425	0.008788	0.069729	1	0.003487	0.00786	0.00934	0.046268	0.016087	0.135649
LTFL	0.376899	0.955236	0.083887	-0.0563	-0.03024	-0.03024	0.960483	0.697931	0.476997	0.065388	0.659749	0.94407	-0.08245	-0.04455	0.899295	-0.3837	-0.08657	0.60938	-0.06054	0.954371	-0.18791	0.003487	1	0.851874	-0.11077	0.006397	-0.06062	0.058351
LTB	0.136127	0.726786	0.007351	-0.04775	-0.01961	-0.01961	0.672432	0.304674	0.187256	0.023969	0.392539	0.894487	0.047132	-0.04168	0.984228	-0.20808	-0.04131	0.132171	-0.04358	0.723368	-0.14056	0.00786	0.851874	1	-0.07276	-0.0009	-0.06652	0.025056
OrdiI/UA	0.273713	-0.07419	0.382212	-0.0951	0.489619	0.489619	-0.11788	-0.01971	0.136731	-0.74223	0.026539	-0.08317	0.26002	0.153129	-0.06224	0.349091	0.160338	-0.09305	0.308123	-0.07078	0.196803	0.00934	-0.11077	-0.07276	1	0.574809	0.333499	0.18495
DiffOrdiI/UA	0.144057	0.001431	0.248193	-0.01367	0.19113	0.19113	0.009519	-0.006	0.04127	-0.50107	0.013599	0.004807	-0.05054	0.020323	0.007765	0.000967	-0.02497	0.002954	0.346136	0.00356	-0.04346	0.046268	0.006397	-0.0009	0.574809	1	0.086997	0.169653
OCF/UA	0.148939	-0.01491	0.080075	-0.08223	0.283646	0.283646	-0.05032	0.081563	0.157872	0.384247	0.165329	-0.03363	0.242055	0.01094	-0.04525	0.274621	0.244446	-0.02213	0.085537	-0.01603	0.35896	0.016087	-0.06062	-0.06652	0.333499	0.086997	1	0.457187
F-score	0.118605	0.06106	0.033571	-0.14437	0.288461	0.288461	0.069147	0.051804	0.145424	0.143866	0.119997	0.042032	0.016809	0.033279	0.047654	0.049674	-0.03478	0.032824	0.070859	0.061331	0.006886	0.135649	0.058351	0.025056	0.18495	0.169653	0.457187	1

다음 [그림 4-2]는 시각화를 나타내는 seaborn 패키지의 하나인 pairplot으로 변수 간 관계를 표현한 것이다. pairplot은 변수 프레임을 인수로 받아 그리드(grid) 형태로 각 변수 열의 조합에 대해 스캐터 플롯을 그린다. 같은 변수가 만나는 대각선 영역에는 해당 변수의 히스토그램을 그리는데, 이를 통해 각 칼럼(column) 별 데이터의 상관관계나 분류적 특정을 확인할 수 있다[16].



[그림 4-2] pairplot으로 나타낸 변수 간 관계 그래프

16) 출처: 데이터 사이언스 스쿨, https://datascienceschool.net/view-notebook/4c2d5ff1caab4b21a708cc662137bc65/

4.3 예측모델 설계

연구설계에 따라 2016년과 2017년 공기업 재무건전성 예측모델을 설계하기 위하여 총 26개 공기업의 2011년터 2017년 7개년 회계연도 데이터를 훈련 데이터(Train data)와 테스트 데이터(Test data)로 나누었다.

연구설계에 따라 2016년 공기업 재무건전성 예측모델을 설계하기 위하여 2011년부터 2015년까지 5개년 재무 데이터를 훈련 데이터로 사용하고, 2016 재무 데이터를 테스트 데이터로 사용하였다. 마찬가지로 2017년 공기업 재무건전성 예측모델을 설계하기 위하여 2012년부터 2016년까지 5개년 재무 데이터를 훈련 데이터로 사용하고, 2017년 재무 데이터를 테스트 데이터로 사용하였다.

이어 4가지 머신러닝 기법을 이용하여 2016년과 2017년 공기업 재무건전성 예측모델을 설계하여 도출된 예측값과 해당 연도 실제값과 비교하여 예측모델의 신뢰성을 검증하였다.

공기업 재무건전성 예측모델 설계하기 위해 사용한 머신러닝 기법은 Random Forest, XGBoost, LightGBM, DNN이다. 각 기법을 이용한 2016년과 2017년 공기업 재무건전성 예측모델 설계 과정과 결과를 차례로 제시하였다.

4.3.1 Random Forest

Random Forest를 이용한 2016년과 2017년 공기업 재무건전성 예측모델 설계 과정과 결과는 다음과 같다. 독립변수와 목표변수가 거의 모두 연속형 변수로 회귀(Regression)식 예측모델을 설계하였다.

먼저 머신러닝 기법을 사용하기 전 지정해야 할 설정들이 있다. 이 중 하나가 하나가 파라미터(parameter)로 기본적으로 함수의 정의 부분에 나열되어 있는 변수들을 의미한다. 매개변수라고도 하는 이 파라미터는 함수의 정의 부분에 포함되어 있는 고유한 특성이다[17].

Random Forest를 이용한 예측모델 설계 시 주요 파라미터(parameters)는 다음과 같이 설정하였다.

```
n_estimators = 1000,
bootstrap = True,
max_features = 'auto',
random_state = 42
```

각 파라미터의 의미는 다음과 같다[18](뮐러 & 가이도, 2017).

- n_estimators = 1000 - 모델을 만들기 위해 생성할 트리(tree)의 개수를 의미한다. 일반적으로 트리의 수가 많을수록 성능이 향상되고, 안정적으로 수행되지만 계산 속도가 느려지는 단점이 있다.
- bootstrap = True - 각 트리가 고유하게 만들어 지도록 데이터의 중복 추출 허용에 대한 선택 여부를 설정하는 것이다.
- max_features = 'auto' - 노드(node) 분할 시 최상의 분할을 위해 고려해야 할 feature(변수)의 수를 의미한다. 이는 각 트리가 얼마나 무작위가 될지를 결정하며, 과적합(over-fitting)을 줄이는 역할을 한다. 'auto'는 feature의 수와 동일하게 설정한 것과 같다. feature의 수는 트리의 각 분기(가지가 나뉘는 부분)에서 모든 특성을 고려하므로 특성 선택에 무작위

17) 출처: 위키백과 https://ko.wikipedia.org/wiki/매개변수_(컴퓨터_프로그래밍)
18) 출처: https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestRegressor.html

성이 들어가지 않지만, 부트스트랩(bootstrap) 샘플링으로 인한 무작위성은 그대로이다.

- random_state = 42 – 난수(random number) 생성기에서 사용하는 시드(seed)로 고정된 값을 사용하여 프로그램 실행 시마다 똑같은 결과를 산출한다.

그 외 나머지 파라미터(parameter)는 RandomForestRegressor 기본값으로 자동 설정된다. 다음은 예측모델 설계에서 훈련 데이터(train data)의 정확도(Accuracy)를 알아보았다. 훈련 데이터로 학습한 모델 훈련 데이터와 특성이 같다면 처음 보는 새로운 데이터가 주어져도 정확히 예측할 거라고 기대한다. 모델이 처음 보는 데이터에 대하여 정확하게 예측할 수 있다면, 이를 훈련 셋(Train set)에서 테스트 셋(Test set)으로 일반화(generalization)되었다고 한다. 우리가 찾으려는 모델은 일반화 성능이 최대가 되는 최적점에 있는 모델이다(뮐러 & 가이도, 2017). Random Forest 이용한 모델 설계에서 훈련 데이터셋(Train Dataset)의 정확도는 〈표 4-7〉과 같다.

〈표 4-7〉 Random Forest 모델 훈련 데이터셋의 정확도

2011-2015년 훈련 데이터셋	2012-2016년 훈련 데이터셋
93.9%	94.4%

2011년에서 2015년까지의 데이터로 구성된 훈련 데이터셋(train dataset)의 정확도는 약 93.6%로 나타났으며, 2012년에서 2016년까지의 데이터로 구성된 훈련 데이터셋의 정확도는 약 94.4%로 나타났다. 데이터셋의 정확도는 파라미터의 설정에 따라 조금씩 달라진다. 때문에 파라미터의 설정을 통해 데이터셋의 정확도를 조정할 수 있다. 하지만 테스트셋의 정확도가 높다고 반드시 좋은 결과가 나오는 것은 아니다.

3장에서 언급한 것처럼, 회귀 예측모델에서 예측에 대한 성능 평가는 주로 오차함수(Error Function) 값으로 확인한다. 본 연구에서는 MAE(Mean Absolute Error, 평균절대오차), MSE(Mean Squared Error, 평균제곱오차),

RMSE(Root Mean Squared Error, 평균제곱근오차), MAPE(Mean Absolute Percentage Error, 평균절대비율오차) 4개의 오차함수 값으로 확인하였다. MAPE의 경우, 다른 평가지표와 달리 %로 표기한다. 예측모델의 오차함수 값이 작을수록 예측 성능이 좋은 것으로 평가한다.

회귀 예측모델에서 주로 가장 많이 쓰이는 성능 평가 지표는 RMSE인데, 수치 예측모델에서는 MAPE도 자주 사용된다. RMSE가 눈금 의존적 오차라면 MAPE는 비율 의존적 오차로, 수치형 데이터의 경우 MAPE의 값이 더 정밀도가 있다고 본다. 이에 본 연구에서의 예측모델 성능 평가 결과는 주로 RMSE와 MAPE를 가지고 설명하였다. Random Forest를 이용한 2016년과 2017년 공기업 재무건전성 예측모델에 대한 성능 평가 결과는 〈표 4-8〉과 같다.

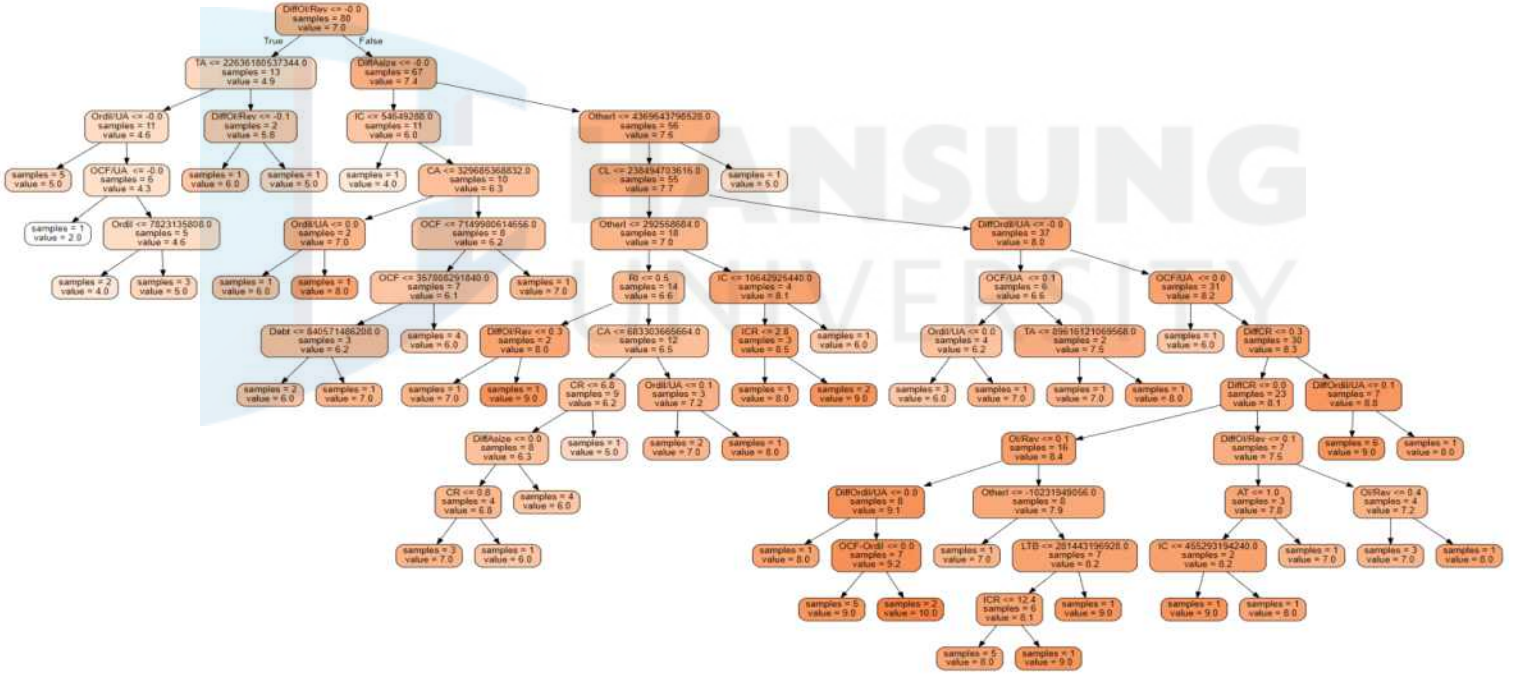
〈표 4-8〉 Random Forest를 이용한 공기업 재무건전성 예측모델의 예측 오차

구분	MAE	MSE	RMSE	MAPE
2016년 예측모델	0.839	0.980	0.989	13.05%
2017년 예측모델	0.932	1.592	1.261	15.79%

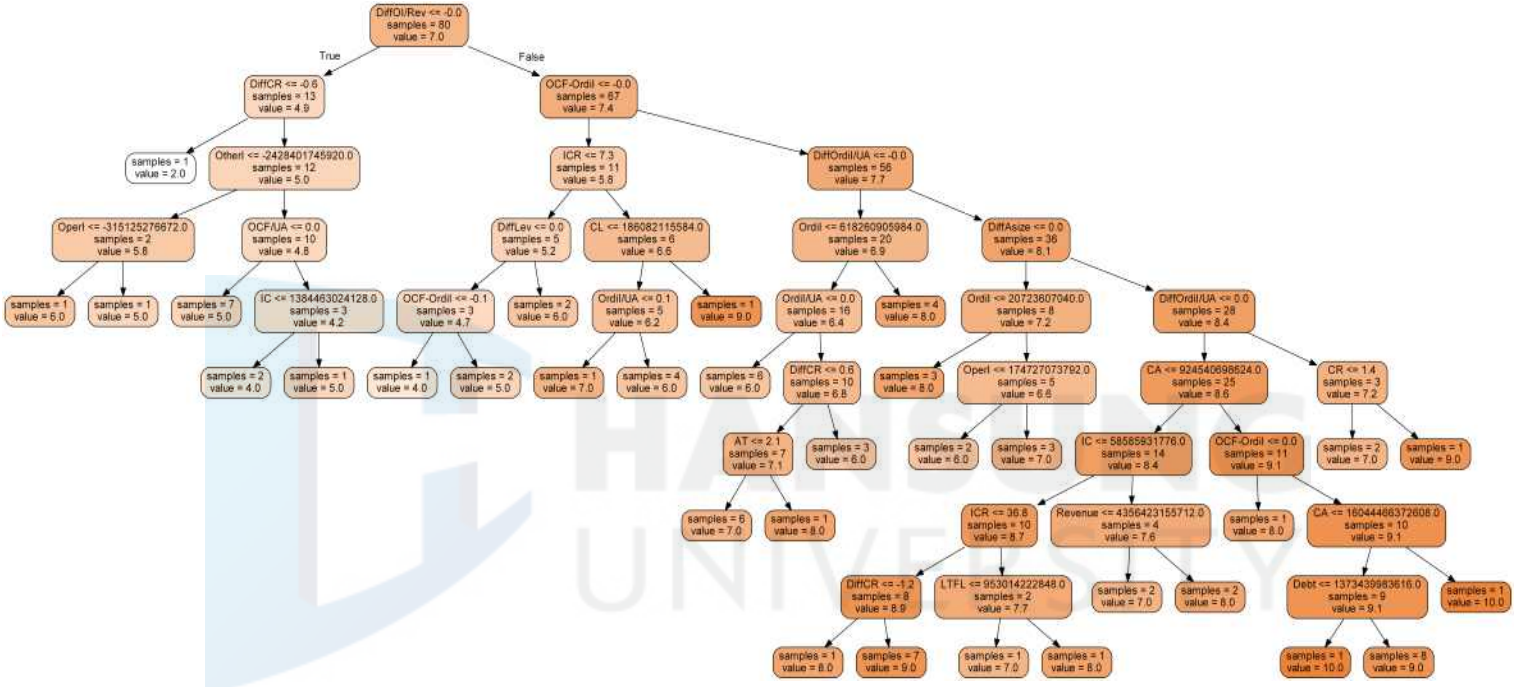
*MAE : Mean Absolute Error, 평균절대오차
*MSE : Mean Squared Error, 평균제곱오차
*RMSE : Root Mean Squared Error, 평균제곱근오차
*MAPE : Mean Absolute Percentage Error, 평균절대비율오차

Random Forest를 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 오차함수 값을 살펴보면, 먼저 2016년 예측모델의 RMSE 값은 0.839이며, 2017년 예측모델의 RMSE 값은 0.939이다. 2016년과 2017년 RMSE 값이 0.093으로 약 0.1정도 차이가 난다. MAPE의 경우, 2016년 예측모델의 MAPE 값은 13.05%, 2017년 예측모델은 15.79%로 2.74%의 차이를 보였다. 즉, 두 가지 지표를 통한 예측 성능 평가 결과, 2016년 예측모델보다 2017년 예측모델의 예측력이 더 낮은 것으로 나타났다. 이와 같은 결과는 2017년 정권 교체의 영향으로 보인다. 2017년 초 탄핵소추로 인한 박근혜 정부가 막을 내리고, 2017년 5월 문재인 정부가 출범하면서 정책 등 사회 전반에 많은 변화 있었다. 이러한 변화는 기존 데이터를 통한 예측의 한계 요인으로 작용할 수 있다.

[그림 4-3], [그림 4-4]는 Random Forest를 이용한 2016년과 2017년 공기업 재무건전성 예측모델에서 단일 변수로 생성된 의사결정트리(Decision Tree) 전체를 시각화하여 나타낸 형태이다. Random Forest를 이루는 각 의사결정트리는 전체 훈련 데이터셋으로부터 일부 무작위 복원 추출된 훈련 데이터와 독립변수에 의해 구축된다. 각 의사결정트리마다 훈련한 데이터셋이 다르기 때문에 구축된 트리 모델로부터 예측된 값이 모두 다르게 나타난다(한은정, 2015).



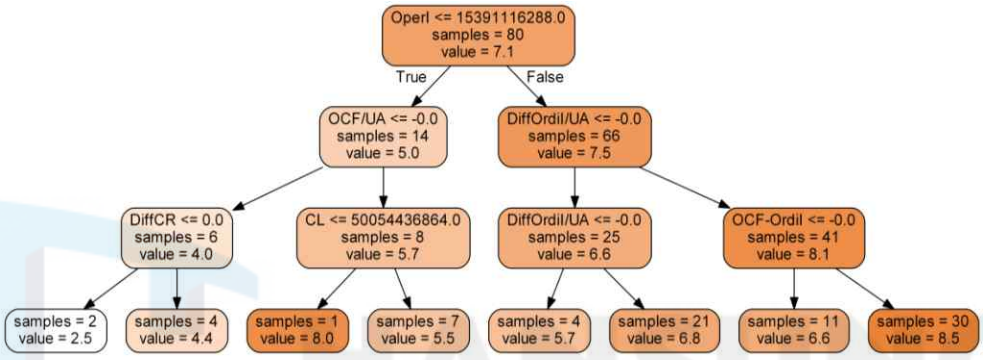
[그림 4-3] Random Forest를 이용한 2016년 공기업 재무건전성 예측모델 단일 전체 결정 트리



[그림 4-4] Random Forest를 이용한 2017년 공기업 재무건전성 예측모델 단일 전체 결정 트리

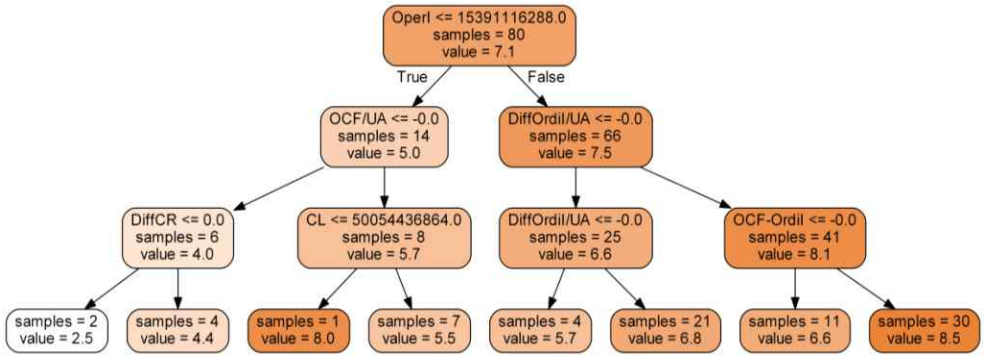
Random Forest는 의사결정트리(Decision Tree) 기법과 달리 배깅(bagging)을 이용한 앙상블 학습 모델로, 변수의 임의 복원(중복허용) 추출을 통해 트리를 생성하고 결합하여 가장 좋은 결과값을 도출한다. 한 번 사용한 변수(트리)의 임의 선택 및 중복 사용을 허용하여 과적합(over-fitting)을 줄이고 일반화 하기가 쉽다.

다음 [그림 4-5], [그림 4-6]은 2016년과 2017년 공기업 재무건전성 예측모델의 의사결정트리의 깊이(depth)를 레벨(level) 3까지 제한하여 시각화하여 나타낸 형태이다. 트리의 깊이는 노드(node)가 나뉘어 갈라지는 분기 시 생성되는 층으로, 가장 상위 노드는 깊이가 0이 된다. [그림 4-5]에서 가장 상위 노드는 OperI(영업이익손실)에 해당하며, Samples는 각 노드에 있는 샘플의 개수를 나타내며, value는 각 노드에 있는 샘플의 예측값을 의미한다.



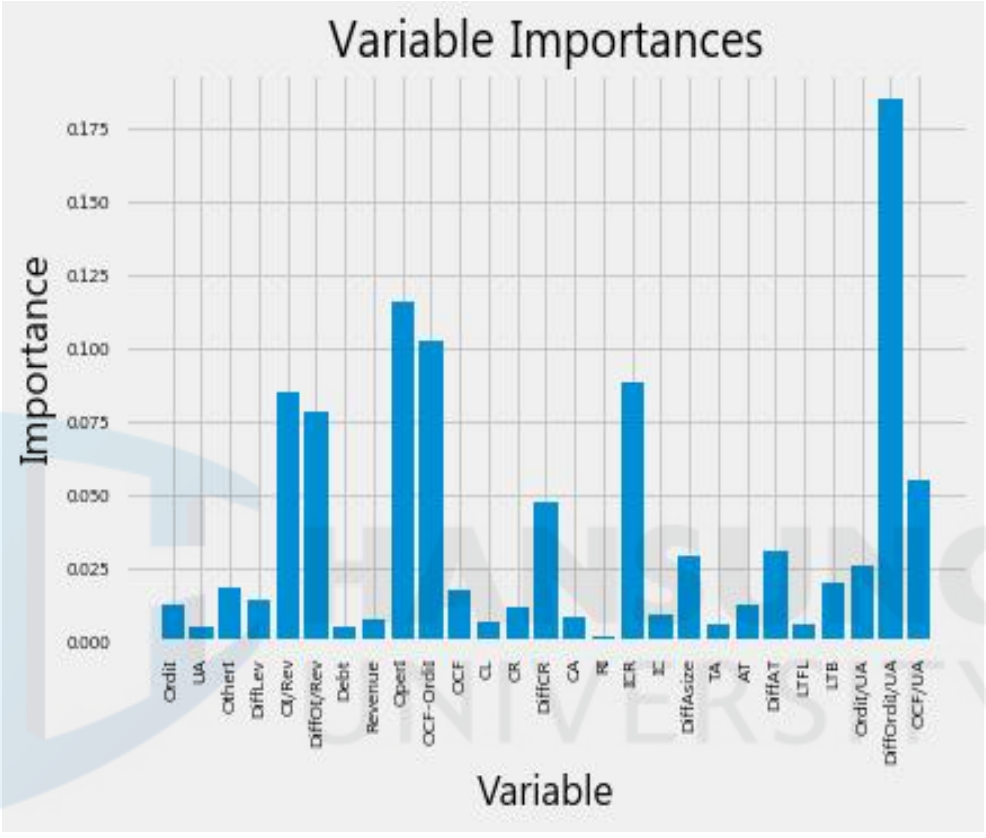
[그림 4-5] Random Forest를 이용한
2016년 공기업 재무건전성 예측모델의 결정 트리 (depth level 3)

Random Forest에서 노드 분기 시 각 노드에서 전체 변수를 대상으로 최선의 테스트를 찾는 것이 아니라 각 노드에서 후보 변수을 랜덤하게 선택한 후 이 후보 변수들 중에서 최선의 테스트를 찾는다(뮐러 & 가이도, 2017).



[그림 4-6] Random Forest를 이용한
2017년 공기업 재무건전성 예측모델의 결정 트리 (depth level 3)

[그림 4-7]은 Random Forest를 이용한 2016년 공기업 재무건전성 예측모델의 변수 중요도(variable[19] importance)를 나타낸 그래프이다.



[그림 4-7] Random Forest를 이용한 2016년 예측모델의 변수 중요도

Random Forest를 이용한 2016년 공기업 재무건전성 예측모델의 변수 중요도(variable importance) 상위 4개 변수와 중요도 값은 다음과 같았다. 변수 중요도가 가장 높은 변수는 총자산경상이익률의변화(DiffOrdiI/UA)로 중요도 값이 0.19였으며, 두 번째 변수는 매출액영업이익률(OI/Rev)로 중요도 값이 0.12였다. 세 번째 변수는 매출영업이익률변화(DiffOI/Rev)로 중요도 값이 0.10, 네 번째 변수는 영업이익손실(OperI)로 중요도 값이 0.09였다.

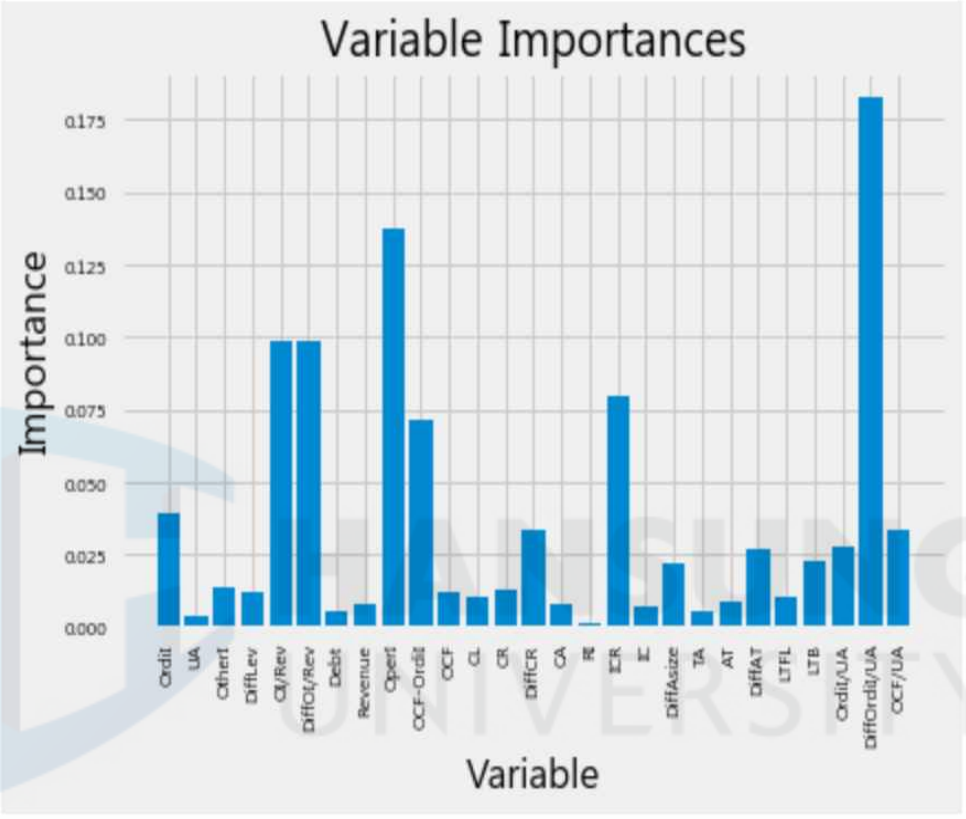
19) 본 연구에서 variable은 머신러닝에서 일반적으로 얘기하는 feature를 의미한다.

Random Forest에서 각 변수의 중요도 값은 각 트리의 변수 중요도를 취합하여 계산한 것이다. 회귀식 예측모델의 경우 어떤 변수를 사용한 노드가 평균제곱오차인 MSE(Mean Square Error)나 평균절대오차인 MAE(Mean Absolute Error)를 사용하여 분산(Variance)을 얼마나 감소시키는지를 계산하며, 각 변수의 상대적 중요도를 살펴볼 수 있다. 예측의 측면에서 독립변수의 중요성을 확인할 수 있으며, 독립변수의 중요도 값이 클수록 목표변수 예측에 중요하다고 해석한다. 일반적으로 Random Forest에서 제공하는 변수 중요도는 하나의 트리에서 제공하는 것보다 더 신뢰할 만하다(뮐러, 가이도, 2017)

중요도가 높은 상위 4개의 변수들만을 가지고 Random Forest를 이용하여 2016년 공기업 재무건전성 예측모델을 다시 설계하였다. 파라미터는 전체 변수를 가지고 Random Forest를 이용한 예측모델 설계 시와 동일하게 설정하였다. 예측모델의 성능 평가는 평균절대비율오차인 MAPE(Mean Absolute Percentage Error) 값으로 측정하였다. MAPE 값이 낮을수록 오차가 적으므로 예측모델의 정확도가 높다고 판정한다. MAPE 값이 오차 비율을 의미하므로, 100-MAPE는 예측 정확도 비율을 의미한다.

상위 4개 변수들로 Random Forest를 이용하여 설계한 2016년 공기업 재무건전성 예측모델의 MAPE 값은 12.35%로 나타났다. MAPE 값이 오차비율이므로, 예측 정확도는 100%에서 12.35%를 뺀 87.65%이다. 전체 변수를 가지고 예측모델을 설계했을 때의 예측 정확도(100-MAPE)가 86.95%이므로, 상위 4개 변수들을 가지고 설계한 예측모델의 정확도가 0.7% 더 높은 것으로 나타났다. 이러한 결과는 상위 4개 변수들만 가지고 2016년 공기업 재무건전성을 예측하는 데 충분히 유용한 지표가 됨을 의미한다.

[그림 4-8]은 Random Forest를 이용한 2017년 공기업 재무건전성 예측 모델의 변수 중요도(variable importance)를 나타낸 그래프이다.



[그림 4-8] Random Forest를 이용한 2017년 예측모델의 변수 중요도

마찬가지로 Random Forest를 이용한 2017년 공기업 재무건전성 예측모델의 변수 중요도(variable importance) 상위 4개 변수와 중요도 점수는 다음과 같았다. 변수 중요도가 가장 높은 변수는 총자산경상이익률의변화(DiffOrdiI/UA)로 중요도 값이 0.18이었으며, 두 번째 변수는 영업이익손실(OperI)로 중요도 값이 0.14였다. 세 번째 변수는 매출액영업이익률(OI/Rev)로 중요도 값이 0.10이었고, 네 번째 변수는 매출영업이익률변화(DiffOI/Rev)로 중요도 값이 0.09였다.

2016년과 동일하게 중요도 상위 4개 변수들만을 가지고 Radom Forest를 이용하여 2017년 공기업 재무건전성 예측모델을 설계하였다. 예측모델의 성능 평가는 평균절대비율오차인 MAPE(Mean Absolute Percentage Error) 값으로 측정하였다. 즉, MAPE 값이 낮을수록 오차비율이 낮아지므로 예측 정확도가 높다고 판정하며, 100-MAPE 값은 예측 정확도에 해당한다.

변수 중요도 상위 4개 변수들로 설계한 2017년 공기업 재무건전성 예측모델의 MAPE 값은 18.10%로 나타났다. MAPE 값이 오차비율이므로, 예측 정확도는 100%에서 18.10%를 뺀 81.90%이다. 전체 변수를 가지고 예측모델을 설계했을 때의 예측 정확도(100-MAPE)가 84.21%이므로, 상위 4개 변수들로 설계한 예측모델의 정확도가 2.22% 더 낮은 것으로 나타났다. 즉, 변수 중요도 상위 4개만으로 만든 2017년 공기업 재무건전성 예측모델의 정확도가 전체 변수로 설계했을 때보다 떨어지는 것으로 나타났다.

이에 2017년 공기업 재무건전성 예측모델의 변수 중요도 개수를 1개에서부터 10개까지 하여 예측모델을 설계해 보았으나, 전체 변수를 가지고 만든 공기업 재무건전성 예측모델의 예측 정확도(100-MAPE) 84.21%보다 모두 낮은 것으로 나타났다. 변수 1개 시 예측 정확도는 80.47%, 변수 2개 시 80.46%, 변수 3개시 81.81%, 변수 5개 시 82.65%, 변수 4개 시 81.90%, 변수 6개 시 83.48%, 변수 7개 시 83.1%, 변수 8개 시 82.67%, 변수 9개 시 83.20%, 변수 10개시 83.36%로 나타났다. 이 중에서 변수 6개로 설계한 예측모델의 애측 정확도(83.48%)이 전체 변수를 가지고 예측모델을 설계했을 때의 예측 정확도(84.21%)와 오차 차이 값이 0.73%로 가장 낮은 것으로 나타났다. 반면 2016년 공기업 재무건전성 예측모델의 경우, 변수 중요도 상위 4개의 변수를 가지고 공기업 재무건전성 예측모델을 설계했을 때의 예측 정확도(87.65%)가 가장 높은 것으로 확인되었다.

Random Forest를 이용한 2016년과 2017년 공기업 재무건전성 예측모델에서 도출된 변수 중요도 상위 4개 변수를 정리하면 〈표 4-9〉와 같다.

〈표 4-9〉 Random Forest를 이용한 예측모델의 변수 중요도 상위 4개

2016년 공기업 재무건전성 예측모델	2017년 공기업 재무건전성 예측모델
1. 총자산경상이익률의변화(DiffOrdiI/UA)	1. 총자산경상이익률의변화(DiffOrdiI/UA)
2. 매출액영업이익률(OI/Rev)	2. 영업이익손실(OperI)
3 .매출영업이익률변화(DiffOI/Rev)	3. 매출액영업이익률(OI/Rev)
4. 영업이익손실(OperI)	4. 매출영업이익률변화(DiffOI/Rev)

〈표 4-9〉에서 보는 바와 같이, Random Forest를 이용한 2016년과 2017년 예측모델의 중요도 상위 4개 변수는 일부 순위에 차이가 있으나 동일한 것으로 나타났다. 중요도 상위 4개 변수는 공기업 재무건전성을 예측할 수 있는 있는 주요 변수에 해당한다. 여기서 중요도 상위 4개 변수인 총자산경상이익율의변화(DiffOrdiI/UA), 매출액영업이익률(OI/Rev), 매출액영업이익률의변화(DiffOI/Rev), 영업이익손실(OperI)은 모두 수익성과 관련이 있는 지표들이다. 또한 중요도 상위 변수 4개와 변수 전체로 Random Forest를 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 성능 평가 결과는 〈표 4-10〉과 같다. 예측 모델의 정확도는 100-MAPE 값으로 측정한다.

〈표 4-10〉 Random Forest를 이용한 예측모델 정확도 (변수 4개 VS 변수 전체)

2016년 공기업 재무건전성 예측모델		2017년 공기업 재무건전성 예측모델	
변수 개수	정확도(%)	변수 개수	정확도(%)
변수 중요도 상위 4개	87.65	변수 중요도 상위 4개	81.90
전체	86.95	전체	84.21

*정확도(%) : 100-MAPE, MAPE(Mean Absolute Percentage Error, 평균절대비율오차)

〈표 4-10〉에서 보는 바와 같이, 중요도 상위 4개 변수를 가지고 설계한 공기업 재무건전성 예측모델의 연도별 예측 정확도는 전체 변수를 가지고 설계한 예측모델의 정확도와 차이를 보였다. 2016년 공기업 재무건전성 예측모델의 경우, 중요도 상위 4개 변수로 설계할 때가 전체 변수로 설계할 때보다 예측 정확도가 높았다. 이는 상위 4개 변수만으로도 2016년 공기업 재무건전성을 예측하는 데 유용한 지표가 됨을 의미한다.

반면 2017년 공기업 재무건전성 예측모델은 중요도 4개 변수로 설계할 때가 전체 변수로 설계할 때보다 예측 정확도가 더 낮았다. 이 경우 중요도가 높은 변수들 중 몇 개의 변수를 선택했을 때 예측의 정확도가 가장 높은지 찾아볼 필요가 있다.

4.3.2 XGBoost(eXtreme Gradient Boosting)

XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 설계 과정과 결과는 다음과 같다.

XGBoost를 이용한 예측모델 설계 시 주요 파라미터(parameters)는 다음과 같이 설정하였다.

```
gamma = 1,
learning_rate = 0.01,
max_depth = 3,
n_estimators = 10000,
subsample = 0.8,
random_state = 42
```

각 하이퍼 파라미터의 의미는 다음과 같다.

- gamma = 1 - 트리의 리프 노드에서 추가 파티션을 만드는 데 필요한 최소 손실 함수를 의미한다. 이것이 커지면, 트리의 깊이가 줄어들고 보수적인 모델이 된다.
- learning rate = 0.01 - 한 번 학습할 때의 학습량을 뜻한다. 학습량이 너무 크거나 작아도 좋지 않다.
- max_depth = 3 - 한 트리의 최대 깊이를 말한다. 3은 기본값이다. 이 값을 늘리면 모델의 복잡도가 커지고, 과적합(over-ftting)이 발생할 수 있다.
- n_estimators = 10000 - 모델을 만들기 위해 생성할 트리의 개수를 의미한다.
- subsample = 0.8 - 훈련 인스턴스의 서브 샘플링 비율을 말한다.
- random_state = 42 - 난수(random number) 생성기에서 사용하는 시드(seed), 고정된 값을 사용하여 프로그램 실행 시마다 똑같은 결과를 산출한다(뮐러 & 가이도, 2017).

XGboost 이용한 모델 설계에서 훈련 데이터셋(Train Dataset)의 정확도는 〈표 4-11〉과 같다.

〈표 4-11〉 XGBoost 모델 훈련 데이터셋의 정확도

2011-2015년 훈련 데이터셋	2012-2016년 훈련 데이터셋
96.3%	96.3%

2011년에서 2015년까지의 데이터로 구성된 훈련 데이터셋(train dataset)과 2012년에서 2016년까지의 데이터로 구성된 훈련 데이터셋(train dataset)의 정확도는 모두 약 96.3%로 나타났다.

Random Forest와 마찬가지로, 예측 모델의 성능 평가는 오차함수(Error Function) 값으로 확인한다. 본 연구에서는 MAE(Mean Absolute Error, 평균절대오차), MSE(Mean Squared Error, 평균제곱오차), RMSE(Root Mean Squared Error, 평균제곱근오차), MAPE(Mean Absolute Percentage Error, 평균절대비율오차) 4개의 오차함수 값으로 확인하였다. MAPE의 경우 다른 지료와 달리 비율 오차로 %로 표기한다. 예측모델의 오차함수 값이 작을수록 예측 성능이 좋은 것으로 평가한다.

XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델에 대한 성능 평가 결과는 〈표 4-12〉와 같다.

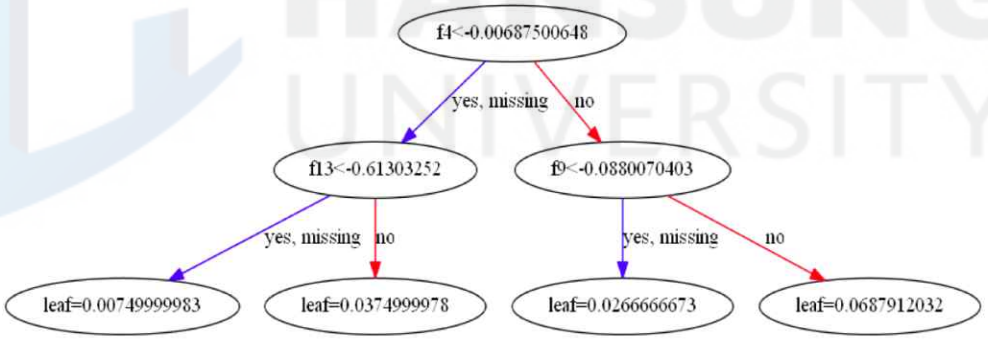
〈표 4-12〉 XGBoost 이용한 공기업 재무건전성 예측모델의 예측 오차

구분	MAE	MSE	RMSE	MAPE
2016년 예측모델	0.800	0.858	0.926	12.54%
2017년 예측모델	0.800	1.121	1.058	13.14%

*MAE : Mean Absolute Error, 평균절대오차
*MSE : Mean Squared Error, 평균제곱오차
*RMSE : Root Mean Squared Error, 평균제곱근오차
*MAPE : Mean Absolute Percentage Error, 평균절대비율오차

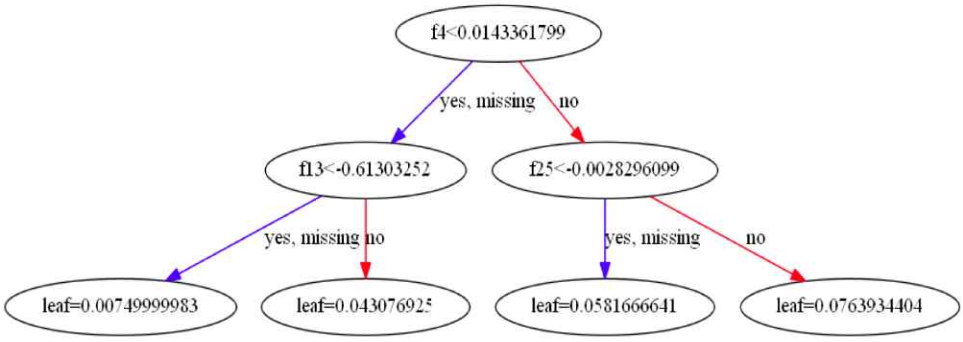
XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 오차 함수 값을 살펴보면, 2016년 예측모델의 RMSE는 0.926, 2017년 예측모델의 RMSE는 1.058이다. 2017년 RMSE 값이 2016년보다 0.074 더 높게 나타났다. MAPE 값 또한 2016년 예측모델이 12.54%, 2017년 예측모델이 13.14%로 2017년 MAPE값이 2016년보다 0.6% 더 높게 나타났다. 오차 값이 낮을수록 예측 정확도가 높은 것이므로, XGBoost를 이용한 2017년 예측모델의 예측정확도가 2016년 예측모델 예측 정확도보다 낮은 것으로 나타났다. 이는 Random Forest와 같은 결과이다. 이러한 결과는 앞서 설명한 바와 같이 2017년 정권 교체로 인한 사회 전반에 걸친 많은 변화 요인이 2017년 공기업 재무건전성 예측에 영향을 준 것으로 보인다.

[그림 4-9], [그림 4-10]는 XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델에서 단일 변수로 생성된 결정트리(Decision Tree) 전체를 시각화하여 나타낸 형태이다.



[그림 4-9] XGboost를 이용한 2016년 예측모델 단일 결정트리

[그림 4-9]에서 노드(node) 내 f4, f13, f9와 숫자는 각각 변수와 변수값에 해당한다. f4, f13, f9는 각각 레버리지변화(DiffLev), 유동비율(CR), 영업이익손실(OperI)을 의미한다. 본 연구에서 트리의 최대 깊이(max_depth)는 기본값인 3으로 설정하여, 하나의 변수로 생성된 트리는 최대 깊이 3을 넘지 않는다.

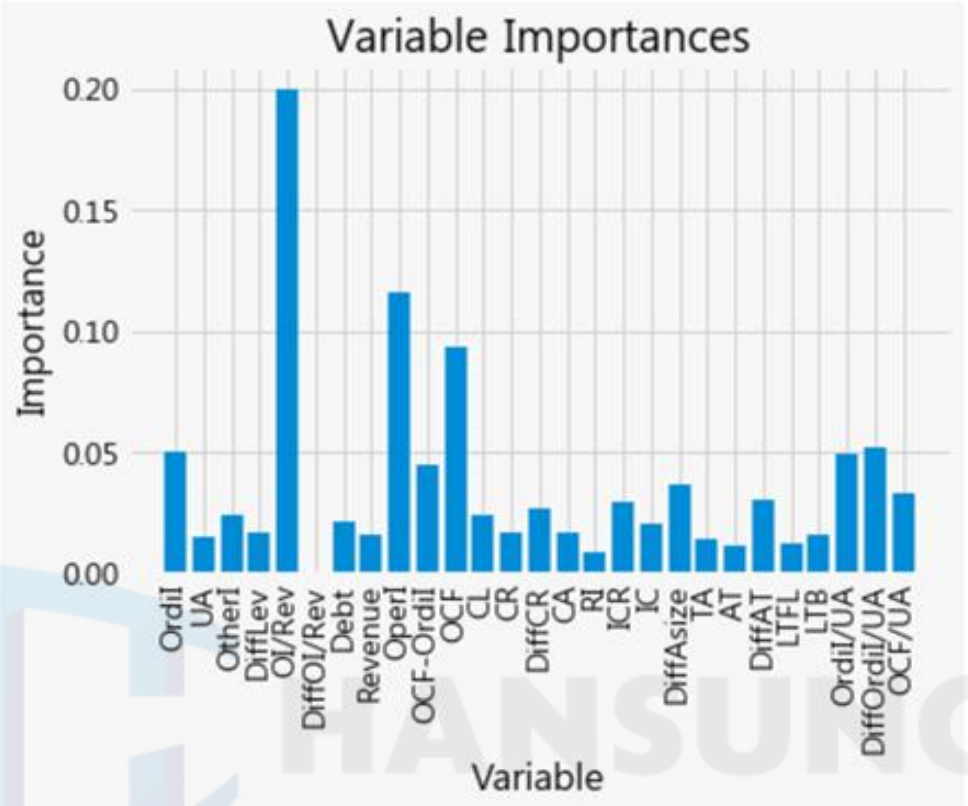


[그림 4-10] XGboost를 이용한 2017년 예측모델 단일 결정트리

[그림 4-10]에서 f4는 레버리지변화(DiffLev), f13은 유동비율(CR), f25는 총자산경산이익율(OrdiI/UA)를 의미한다.

XGBoost의 경우, 트리를 만들 때 CART (Classification And Regression Tree)라 불리는 앙상블 모델을 사용한다. 이후 트리 부스팅을 사용하여, 각 분류기 간 비중(weight)을 최적화한다(Chen & Cusestrin, 2016). XGboost는 이전 트리에서 얻은 정보를 가지고 다음 트리를 생성하는데 활용하여 약한 개체에 가중치를 부여하여 강한 개체로 변환시키면서 오차를 교정해 나간다. 또한 파라미터로 지정한 최대 깊이(max_depth)까지 진행한 후 손실 함수(loss function)에서 개선이 일정 수준에 미치지 못할 경우, 역방향으로 가지치기 과정을 진행한다. 이를 통해 트리 간 모델을 상호보완하여 오차를 줄여 트리 기반 모델에서 발생할 수 있는 과적합(over-fitting) 문제를 해결한다(하지은, 2017).

[그림 4-11]는 XGBoost를 이용하여 만든 2016년 공기업 재무건전성 예측모델의 변수 중요도(variable importance)를 나타낸 그래프이다. XGBoost 또한 Random Forest와 마찬가지로 변수 중요도(variable importance)를 제공하는데 각 트리의 변수 중요도를 취합하여 계산한 것이다. 예측의 측면에서 독립변수의 중요성을 확인할 수 있으며, 독립변수의 중요도 값이 클수록 해당 변수의 예측에 중요하다고 볼 수 있다.



[그림 4-11] XGBoost를 이용한 2016년 예측모델의 변수 중요도

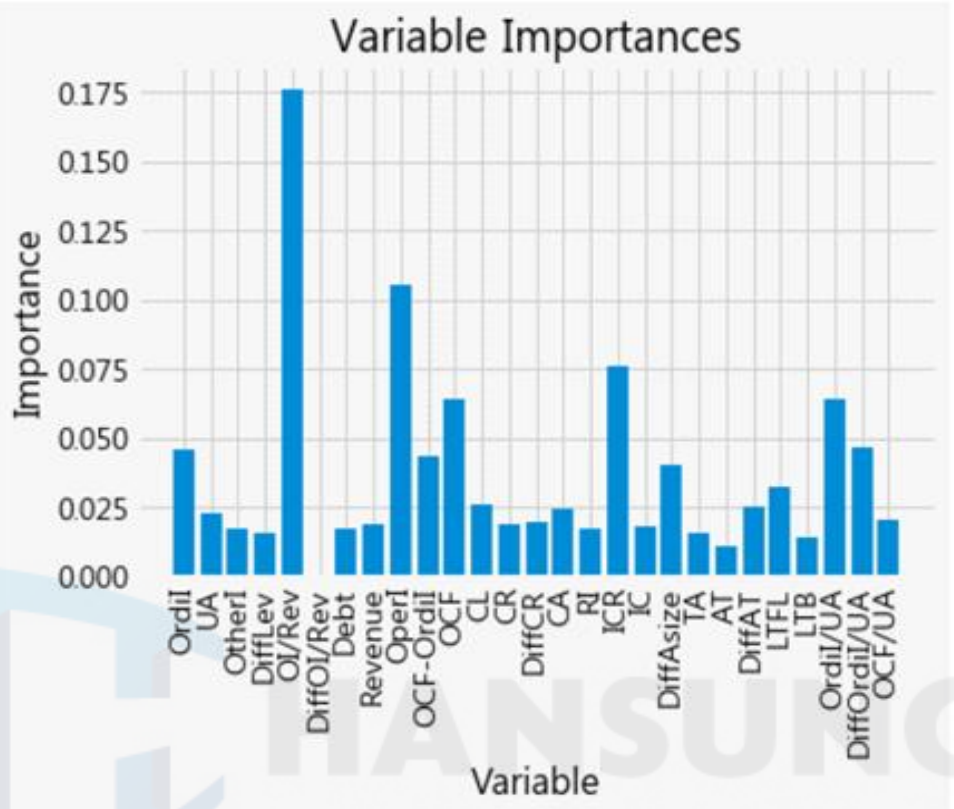
XGBoost를 이용한 2016년 공기업 재무건전성 예측모델의 중요도 상위 4개 변수와 중요도 점수는 다음과 같았다. 중요도가 가장 높은 변수는 매출액영업이익률(OI/Rev)로 0.20이었으며, 두 번째 변수는 영업이익손실(OperI)로 0.11이었다. 세 번째 변수는 영업활동현금흐름(OCF)으로 0.09였고, 네 번째 변수가 총자산경상이익률의변화(DiffOrdiI/UA)로 0.09였다.

마찬가지로, 중요도가 상위 4개의 변수들만을 가지고 XGboost를 이용하여 2016년 공기업 재무건전성 예측모델을 다시 설계하였다. 파라미터는 전체 변수를 가지고 XGboost를 이용한 예측모델 설계 시와 동일하게 설정하였다. 예측모델의 성능 평가는 평균절대비율오차인 MAPE(Mean Absolute Percentage Error) 값으로 측정하였다. 즉, MAPE 값이 낮을수록 오차비율이 낮아지므로 예측 정확도가 높다고 판정하며, 100-MAPE 값은 예측 정확도에 해당한다.

중요도 상위 4개 변수들로 XGboost를 이용하여 설계한 2016년 공기업 재무건전성 예측모델의 MAPE 값은 15.50%로 나타났다. MAPE 값이 오차비율이므로, 예측 정확도는 100%에서 15.50%를 뺀 84.50%이다. 전체 변수를 가지고 예측모델을 설계했을 때의 예측 정확도(100-MAPE)가 87.46%로, 중요도 상위 4개 변수들로 설계한 예측모델의 정확도가 2.96% 더 낮게 나타났다. 즉, XGBoost를 이용한 2016년 공기업 재무건전성 예측모델에서 중요도 상위 4개만으로 설계했을 때가 전체 변수로 설계했을 때보다 예측 정확도가 더 떨어지는 것으로 나타났다.

이에 2017년 공기업 재무건전성 예측모델의 중요도 개수를 상위 1개에서 상위 10개까지의 변수로 예측모델을 설계해 보았으나, 전체 변수를 가지고 XGboost를 이용한 공기업 재무건전성 예측모델의 예측 정확도(87.46%)보다 모두 낮은 것으로 나타났다.

[그림 4-12]는 XGBoost를 이용하여 만든 2017년 공기업 재무건전성 예측모델의 변수 중요도(variable importance)를 나타낸 그래프이다.



[그림 4-12] XGBoost를 이용한 2017년 예측모델의 변수 중요도

XGBoost를 이용한 2017년 공기업 재무건전성 예측모델의 중요도 상위 4개 변수와 중요도 값은 다음과 같았다. 중요도가 가장 높은 변수는 매출액영업이익률(OI/Rev)로 중요도 값이 0.18이었으며, 두 번째 변수와 세 번째 변수는 각각 영업이익손실(OperI)로, 영업활동현금흐름(OCF)으로 중요도 값이 0.08로 같았다. 네 번째 변수는 이자보상배율(ICR)로 중요도 값이 0.07이었다.

중요도 상위 4개의 변수만을 가지고 XBoost를 이용하여 2017년 공기업 재무건전성 예측모델을 다시 설계하였다. 예측모델의 성능 평가는 평균절대비율오차인 MAPE(Mean Absolute Percentage Error) 값으로 측정하였다. MAPE 값이 낮을수록 오차비율이 낮아지므로 예측 정확도가 높다고 판정하며, 100-MAPE 값은 예측 정확도에 해당한다.

중요도 상위 4개 변수로 XGboost를 이용한 2017년 공기업 재무건전성 예측모델의 MAPE 값은 21.89%로 나타났다. MAPE 값이 오차비율이므로, 예측 정확도는 100%에서 21.89%를 뺀 78.11%이다. 전체 변수를 가지고 예측모델을 설계했을 때의 예측 정확도(100-MAPE)가 86.85%이므로, 상위 4개 변수로 설계한 예측모델의 예측 정확도가 전체 변수로 설계한 예측모델의 예측 정확도보다 8.74% 더 낮은 것으로 나타났다.

이에 중요도 개수를 상위 1개에서부터 상위 10개를 가지고 차례로 XGBoost를 이용한 공기업 재무건전성 예측모델을 설계해 보았다. 그 결과 중요도 상위 1개일 때 예측 정확도(100-MAPE)가 79.67%, 상위 2개일 때 82.15%, 상위 3개일 때 78.87%, 상위 4개일 때 78.11%, 상위 5개일 때 77.60%, 상위 6개일 때 83.41%, 상위 7개일 때 83.20%, 상위 8개일 때 84.35%, 상위 9개일 때 84.96%%, 상위 10개일 때 85.29%로 나타났다. 상위 10개 일때가 가장 높아 추가로 상위 11개일 때와 상위 12개 일때의 예측 정확도를 조사한 결과, 상위 11개일 때는 86.25%, 상위 12개일 때 85.2%였다. 이는 전체 변수를 가지고 예측모델을 설계했을 때의 예측 정확도(86.65%) 보다 모두 다 낮은 것으로 나타났다.

XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델에서 도출된 변수 중요도 상위 4개 변수를 정리하면 〈표 4-13〉과 같다.

〈표 4-13〉 XGBoost를 이용한 예측모델의 변수 중요도 상위 4개

2016년 공기업 재무건전성 예측모델	2017년 공기업 재무건전성 예측모델
1. 매출액영업이익률(OI/Rev)	1. 매출액영업이익률(OI/Rev)
2. 영업이익손실(OperI)	2. 영업이익손실(OperI)
3. 영업활동현금흐름(OCF)	3. 영업활동현금흐름(OCF)
4. 총자산경상이익률의변화(DiffOrdiI/UA)	4. 이자보상배율 (ICR)

〈표 4-13〉에서 보는 바와 같이, XGBoost를 이용하여 설계한 2016년과 2017년 예측모델의 중요도 상위 3개 변수가 같고, 나머지 1개 변수가 다르게

나타났다. 연도별 중요도 상위 4개 변수 중 매출액영업이익률(OI/Rev), 영업이익손실(OperI), 영업활동현금흐름(OCF), 총자산경상이익율의변화(DiffOrdiI/UA)는 수익성과 관련된 지표이며, 이자보상배율(ICR)은 재무구조 지표와 연관이 있다. 앞서 설명한 바와 같이 예측모델에서 중요도가 높은 변수는 목표변수를 예측할 수 있는 주요 변수가 될 수 있다.

중요도 상위 4개를 가지고 XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델 성능 평가 결과는 〈표 4-14〉와 같다. 예측 모델의 정확도는 100-MAPE 값으로 측정하였다.

〈표 4-14〉 XGBoost를 이용한 예측모델 정확도 (변수 4개 VS 변수 전체)

2016년 공기업 재무건전성 예측모델		2017년 공기업 재무건전성 예측모델	
변수 개수	정확도(%)	변수 개수	정확도(%)
중요도 상위 4개	84.50	중요도 상위 4개	78.11
전체	87.46	전체	86.85

*정확도(%) : 100-MAPE, MAPE(Mean Absolute Percentage Error, 평균절대비율오차)

〈표 4-14〉에서 보는 바와 같이, 중요도 상위 4개 변수만 가지고 XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 예측 정확도는 전체 변수로 설계 시보다 모두 낮게 나타났다. 이는 상위 4개 변수만으로 2017년 공기업 재무건전성을 예측할 수 있지만, 전체 변수로 예측했을 때보다 예측의 정확도를 떨어질 수 있음을 말해준다. 이 경우 중요도가 높은 변수들 몇 개를 선택했을 때 예측의 정확도가 가장 높은지 찾아볼 필요가 있다.

4.3.3 LightGBM (Light Gradient Boosting Machine)

LightGBM을 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 설계 과정과 결과는 다음과 같다.

LightGBM도 앞서 설정한 Random Forest, XGBoost와 같이 회귀(Regression)식 형태로 예측모델을 설계하였다.

LightGBM 모델링에서는 하이퍼파라미터(hyperparameter)를 설정한다. 머신러닝에서 하이퍼파라미터란 학습 프로세스를 시작하기 전에 설정하는 파라미터 값이다. 반면, 다른 파라미터의 값들은 훈련을 통해 도출된다[20]. 모델 학습 및 테스트에 필요한 시간은 하이퍼파라미터의 선택에 따라 달라질 수 있다(Claesen, Marc & Moor, 2015). LightGBM을 이용한 예측모델 설계 시 하이퍼파라미터(hyperparameter)는 다음과 같이 설정하였다.

‘task’: 'train',
‘boosting_type: 'gbdt',
‘objective: 'regression',
‘metric’: 'rmse',
‘is_training_metric’: True,
‘numleaves’ : 20
‘learning_rate’: 0.05,
‘feature_fraction’: 0.9,
‘bagging_fraction’: 0.8,
‘bagging_freq’: 5,
‘min_data_in_leaf’ : 4
‘verbose’ : 10

20) 출처: WIKIPEDIA. https://en.wikipedia.org/wiki/Hyperparameter_(machine_learning)

각 하이퍼 파라미터의 의미는 다음과 같다[21].

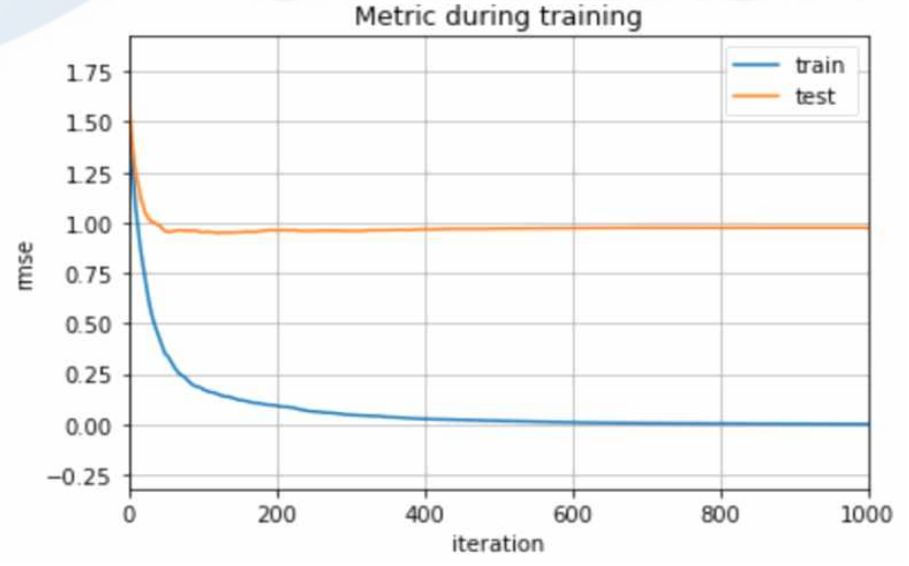
- task: train – train이 기본값이다.
- boosting_type : gbdt – 부스팅(boosting) 형태이다. gbdt는 전통적인 그레디언트 부스팅 결정 트리(Gradient Boosting Decision Tree: GBDT)를 의미한다.
- objective: regression – regression이 기본값이다.
- metric: rmse – 모델 측정 지표로 RMSE(Root Mean Squared Error, 평균제곱근오차)로 설정하였다.
- is_training_metric: True – 모델 훈련 시 측정 지표 사용 여부를 말한다.
- num_leaves : 20 – 트리 모델의 복잡성을 제어하는 주요 매개변수이다. 이론적으로, depth-wise 트리처럼 잎을 얻도록 설정할 수 있지만, 이러한 단순 변환은 실제로 좋지 않다. 그 이유는 하나의 leaf-wise tree가 일반적으로 고정된 수의 잎에 대해 하나의 depth-wisee 트리보다 훨씬 더 깊기 때문이다. 제한되지 않은 깊이는 과적합을 유발할 수 있다. 따라서 num_leaves를 조정하려고 할 때 max_deapth의 제곱보다 작게해야 한다. 예를 들어, max_depth가 7인 depth-wise 트리는 좋은 정학도를 얻을 수 있지만, num_leaves를 127까지 설정하면 과적합을 일으킬 수 있다. num_leaves를 70에서 80까지 설정할 때 depth-wise보다 더 좋은 정확도를 얻을 수 있다.
- learning_rate: 0.05 – 한 번 학습할 때의 학습량을 의미한다.
- feature_fraction: 0.9 – 칼럼(column) 샘플링을 통해 각각의 다양성을 높임. 1이 기본값이지만, 대개 0.7~0.9 세팅이 일반적이다.
- bagging_fraction: 0.8 – 각 반복에 사용될 데이터의 비율을 지정하며 일반적으로 훈련 속도를 높이고 과적합(over-fitting)을 피하는 데 사용한다.
- bagging_freq: 5 – iteration 몇 번째에 해당하는 데이터를 업데이트 할 것인지를 설정하는 것이다.
- min_data_in_leaf : 4 – 잎이 많은 나무에서 과적합을 방지하는 매우 중요한 매개 변수이다. 최적값은 훈련 샘플 수에 따라 다르다.
- verbose: 10 – 학습 중 출력되는 문구를 설정하는 것이다.

21) 출처: lightgbm.readthedocs.io/en/latest/Parameters-Tuning.html

이 외에 설정하지 않은 파라미터는 기본값으로 자동 설정된다.

주요 파라미터 설정 후 학습을 진행할 훈련 모델을 세팅하였다. 학습할 데이터와 훈련 중 평가할 데이터를 설정하고, 학습 진행 시 부스팅(Boosting) 회수를 1,000번으로 설정하였다. 검증 데이터셋에 있는 모든 항목에 대한 평가 결과를 저장하는데 사용되는 evals_result값은 evals_result로 설정하였다. verbose_eval은 10으로 설정하였는데, 이것은 적어도 하나의 검증 데이터가 요구된다. 그 값이 True인 경우 평가 데이터셋에 있는 eval 메트릭이 각 부스팅 단계에서 표기된다. 만약 정수이면 평가 데이터셋의 eval 메트릭이 모든 verbose_eval 부스팅 단계에서 표기된다. 마지막 부스팅 단계 또는 early_stopping_rounds 사용으로 발견된 부스팅 단계로 표기된다. 마지막으로 early_stopping_rounds 파라미터를 10,000번 설정 후 학습을 진행하여 2016년과 2017년 공기업 재무건전성 예측모델 설계 시의 유효한 오차(error) 값을 찾고자 하였다.

모델 학습 결과, 학습에 대한 유효한 오차 값을 찾지 못하였다. LightGBM을 이용한 2016년 공기업 재무건전성 예측모델 학습량에 따른 RMSE(Root Mean Squared Error, 평균제곱근오차) 변화 그래프는 [그림 4-13]과 같다.

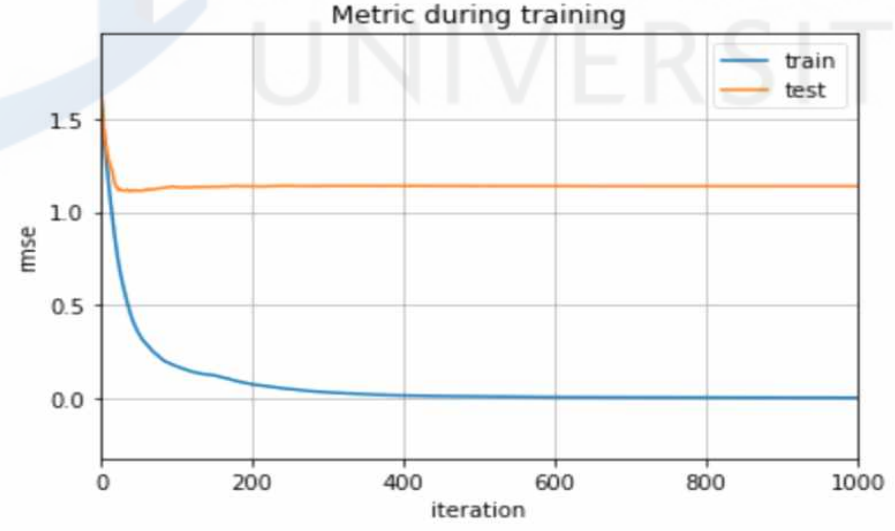


[그림 4-13] LightGBM을 이용한 2016년 예측모델의 학습량에 따른 RMSE

[그림 4-13]에서 보는 바와 같이 훈련 데이터(train data)의 학습량이 늘어날수록 평균제곱근오차인 RMSE(Root Mean Squared Error)값이 계속 낮아져 400회 이상 학습 이후 거의 0에 가깝게 수렴하는 상황이다. 테스트 데이터(test data)도 RMSE값이 낮아져 1.0 부근에서 값이 높아졌다가 낮아지기를 반복했다. 훈련 데이터셋의 정확도가 너무 좋은 경우에도 과적합이라고 본다.

이러한 과적합을 해결하는 가장 기본적인 방법은 학습 데이터를 더 많이 수집하는 것이다. 학습 데이터가 많을수록 실제 데이터를 잘 반영할 수 있기 때문에 과적합 현상이 발생하지 않게 된다. 두 번째는 변수의 개수를 줄이는 방법이 있다. 변수의 개수를 줄여 데스트해 보는 것인데, 변수들 사이에 연관성이 있는 경우도 있어 효과가 발생할 수도 있다. 마지막으로 정규화(Regularization)하는 방법이 있다[22].

LightGBM을 이용한 2017년 공기업 재무건정성 예측모델의 학습량에 따른 RMSE(Root Mean Squared Error, 평균제곱근오차) 변화 그래프는 [그림 4-14]와 같다.



[그림 4-14] LightGBM을 이용한 2017년 예측모델의 학습량에 따른 RMSE

22) 출처: https://docs.aws.amazon.com/ko_kr/machine-learning/latest/dg/model-fit-underfitting-vs-overfitting.html

[그림 4-16]에서 보는 바와 같이 2016년 예측모델과 같은 형태를 보였다. 훈련 데이터(train data)는 학습량이 늘어날수록 RMSE(Root Mean Squared Error)값이 계속 낮아져 거의 0에 가깝게 수렴하는 상황이다. 반면 테스트 데이터(test data)는 RMSE값이 낮아져 1.2 정도 부근에서 값이 올랐다가 내려갔다를 반복한다. 마찬가지로 과적합(over-fitting)이라고 본다.

모델 설계 결과, 본 연구에서 사용된 데이터셋의 규모로는 LighgGBM을 이용하여 공기업 재무건전성 예측모델을 설계하는 데 적합하지 않은 것으로 나타났다.

4.3.4 DNN(Deep Neural Network, 심층신경망)

DNN(Deep Neural Network, 심층신경망) 기법을 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 설계 과정과 결과는 다음과 같다.

먼저 2016년과 2017년 공기업 재무건전성 예측모델 설계에 맞게 훈련 데이터셋(train dataset)과 데스트 데이터셋(test dataset)을 설정하였다. DNN 모델링 하기 전 reshape 함수를 이용하여 목표변수의 배열을 행에서 열 형태로 바꾸었다. reshape 함수는 데이터를 변경하지 않고 배열에 새로운 모양을 부여한다[23]. 이어 스케일링(Scaling)을 적용하였다. 스케일링은 데이터셋에 적용되는 전처리 과정으로 모든 데이터에 선형 변환을 적용하여 전체 데이터의 분포를 평균 0, 분산 1이 되도록 만드는 과정이다. 스케일링은 데이터의 오버플로우(overflow)나 언더플로우(underflow)를 방지하고 독립변수의 공분산 행렬 조건 수를 감소시켜 최적화 과정에서 안정성 및 수렴 속도를 향상시킨다[24].

본 연구에서는 StandardScaler를 통해 데이터의 평균이 0, 표준편차가가 1이 되도록 변환하였다. 훈련 데이터를 입력으로 하여 fit 함수를 사용하여 분포 모수를 객체내에 저장한 후 transform 함수를 실행하여 훈련 데이터와 테스트 데이터를 각각 변환하였다. StandardScaler와 transform을 통한 데이터 설정 형태는 다음과 같다.

```
X_scaler = StandardScaler().fit(X_train)
X_train_scaled = X_scaler.transform(X_train)
X_test_scaled = X_scaler.transform(X_test)
```

딥러닝(Deep Learning)은 퍼셉트론(perceptron) 위에 숨겨진 퍼셉트론 층(layer)을 차곡차곡 추가하는 형태이다. 이 층들이 케라스(Keras)에서는 Sequential 함수를 통해 쉽게 구현된다. Sequential 함수를 model로 선언해 놓고 model.add라는 라인을 추가하면 새로운 층이 만들어진다. 맨 마지막 층은 결과를 출력하는

23) 출처: SciPy.org, https://docs.scipy.org/doc/numpy/reference/generated/numpy.reshape.html
24) 출처: 데이터 사이언스 스쿨, https://datascienceschool.net/view-notebook/f43be7d6515b48c0beb909826993c856/

‘출력층(output layer)’이 되며, 나머지는 모두 ‘은닉층(hidden layer)’이 된다. 첫 번째 층의 경우, ‘입력층(input layer)’이면서 ‘은닉층(hidden layer)’의 역할을 겸한다. 각각의 층은 Dense 함수를 통해 구체적으로 그 구조가 결정된다(조태호, 2017)

이에 케라스(Keras) 패키지를 이용하여 DNN(Deep Neural Network)을 이용한 2016년 Sequential 모델을 만든 형태는 다음과 같다.

```
umber_inputs = 27
number_hidden = 3
number_classes = 1
model.add(Dense(units = number_hidden,  input_dim=number_inputs,
                activation ='relu')
model.add(Dense(units = 7, activation ='relu')
model.add(Dense(units = number_classes, activation = 'linear')
```

Sequential 모델은 총 3개의 층으로 설정하였다. 먼저 Dense 함수를 이용하여 DNN(Deep Neural Netwrok) 은닉층(hidden layer)의 노드(node) 3개를 만들고, input_dim을 통해 입력 데이터의 개수는 독립변수의 수에 해당하므로 27개를 설정하였다. 이는 27개의 변수를 받아 은닉층(hidden layer)의 3개 노드로 보낸다는 의미이다. 은닉층(hidden layer)의 각 노드는 27개의 입력 값으로부터 임의의 가중치를 가지고 각 노드로 전송되어 활성화(activation) 함수를 만난다. 이어 활성화 함수를 거친 결과 값은 출력층(output lyaer)으로 전달된다. 첫 번째 층의 활성화 함수는 relu로 설정하였다. relu는 입력값이 0을 넘으면 그 입력값을 그대로 출력하고, 입력값이 0 이하이면 0을 출력하는 함수를 말한다.

두 번째 층은 은닉층(hidden layer)의 노드를 7개 만들고, 활성화 함수는 동일하게 relu로 설정하였다. 마지막 층인 세 번째 층은 출력층(output layer)으로 출력값을 하나로 정해서 보여줘야 하므로, 출력층(output layer)의 노드는 1개이다. 마찬가지로, 이 노드에서 입력 값 역시 활성화 함수를 거쳐 최종 출력 값으로 나와야 한다. 회귀(regression)식 모델에서는 선형(linear) 함수가

기본 설정값이므로, 본 연구에서는 선형 함수를 활성화 함수로 설정하였다.

이렇게 만들어진 모델은 summary 함수로 모델 내부의 층 리스트를 살펴볼 수 있다. 그 형태를 표로 나타내면 〈표 4-15〉와 같다.

〈표 4-15〉 DNN을 이용한 2016년 Sequential 모델의 내부 층 리스트

Layer (type)	Output shape	Param #
dense (Dense)	(None, 3)	84
dense_1 (Dense)	(None, 7)	28
dense_2 (Dense)	(None, 1)	8

Total params: 120
Trainable params: 120
Non-trainable params: 0

다음으로 모델을 학습하기 전 컴파일(compile) 함수를 통해 모델이 효과적으로 구현될 수 있게 여러 가지 환경을 설정해 준다. 설정한 형태는 다음과 같다.

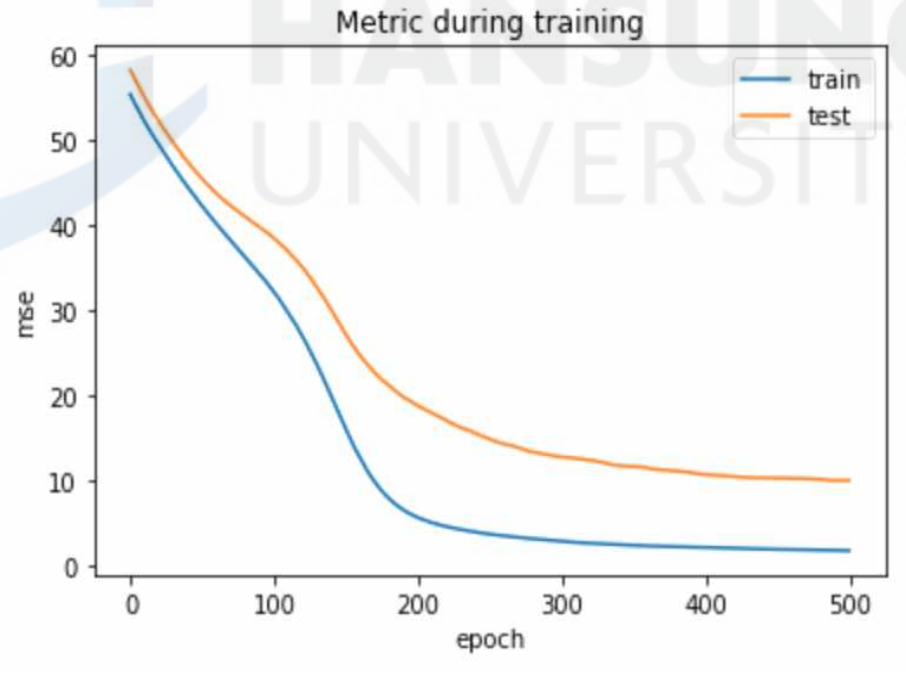
```
model.compile(loss = 'mean_squared_error',
              optimizer = 'adam' , metrics =['mse'])
```

먼저 오차(error) 함수를 설정해주는데, 본 연구에서는 평균제곱오차인 MSE(Mean Squared Error)를 사용하였다. 이어 최적화(optimizer)를 위해 아담(adam)을 사용하였다. 다음으로 측정 지표(metrics) 함수는 모델이 컴파일(compile)될 때 모델 수행 결과를 나타낸다. 측정 지표(metrics) 함수는 정확도를 측정하기 위해 사용되는 테스트 샘플을 학습 과정에서 제외시킴으로써 과적합(over-fitting)을 방지한다(조태호, 2017).

다음으로 fit 함수를 이용하여 모델을 학습시켰다. 먼저 훈련 데이터 X_train_scaled, y_train)를 설정했다. 본 연구에서는 독립변수의 데이터를 스케일링(scaling)을 통해 평균이 0, 표준편차가 1이 되도록 만들었다.

학습 프로세스가 모든 샘플에 대해 한 번 실행되는 것을 1 epoch라고 한다. epoch를 1,000으로 놓으면 각 샘플이 처음부터 끝까지 100번 재사용될 때까지 실행을 반복하라는 뜻이다(조태호, 2017). 본 연구에서는 유효한 값을 찾기 위해 숫자를 계속 바꾸어 입력하였다. shuffle은 주어진 배열의 순서를 뒤섞는 함수를 의미한다(최용, 2019) verbose는 학습 중 출력되는 문구를 설정하는 것이다. 주피터 노트북(Jupyter Notebook) 사용 시에는 verbose를 2로 설정하여 진행 막대(progress bar)가 나오지 않도록 설정한다[25].

모델 학습 결과, 학습에 대한 유효한 값을 찾지 못하였다. 이는 본 연구에서 사용한 데이터 수가 작음으로 인해 나타난 과적합(over-fitting) 현상으로 보인다. epoch, 즉 학습 프로세스가 모든 샘플에 대해 한 번 실행되는 회수에 따른 오차함수 MSE(Mean Squared Error, 평균절대오차)의 변화 그래프는 [그림 4-15]와 같다.



[그림 4-15] DNN을 이용한 2016년 예측모델의 epoch에 따른 MSE

25) 출처: 데이터 사이언스 스쿨. https://datascienceschool.net/view-notebook/51e147088d474fe1bf32e394394eaea7/

[그림 4-15]에서 보는 바와 같이 훈련 데이터셋(train dataset)과 테스트 데이터셋(test dataset) 모두 epoch(학습 프로세스가 모든 샘플에 대해 한 번 실행되는 회수)가 늘어날수록 MSE(Mean Squared Error, 평균절대오차)가 급격히 낮아지더니 둘 다 일정한 값에 일정한 값에 이르는 것처럼 보인다. 하지만, 이미 훈련 데이터셋의 MSE 값이 거의 0에 가까이 수렴하고 있어 이는 과적합(over-fitting) 상태로 본다. 즉, 훈련 데이터셋 안에서는 일정 수준 이상의 예측 정확도를 보이지만, 실제 새로운 데이터에 적용하면 잘 맞지 않는다는 것이다.

신경망 학습에서 학습시켜야 할 모델의 파라미터들(parameters)에 비해 데이터의 수가 충분하지 않은 경우 과적합 문제가 발생한다(Srivastava et al., 2014). 때문에 과적합을 방지하기 위해 L1과 L2 정규화, 드롭아웃(dropout), data augmentation, early stopping 등의 기법을 사용하고, 네트워크 사이즈를 줄이거나 학습률(learning rate)을 낮추는 방법이 적용되어 왔다(Nowlan & Hinton, 1992; Srivastava et al., 2014; Goodfellow, Bengio & Courville, 2016).

하지만 이러한 방법들이 과적합을 방지할 수는 있지만, 일반화 성능을 향상시킬 수는 없다. 네트워크의 사이즈를 줄이는 것 또한 과적합을 방지할 수 있지만 일반화 성능을 향상시킬 수 없다(백유진, 2019). 과적합 문제를 해결하는 가장 일반적인 방법은 해결책은 학습 데이터를 증가시키는 것이다. 과적합 문제는 신경망 연구에 중요한 부분으로 현재도 이 문제를 해결하기 위한 많은 연구가 진행되고 있다(윤운규, 2018).

DNN(Deep Neural Network)을 이용하여 2017년 Sequential 모델 또한 만들었다. 2016년 모델과 마찬가지로 2017년 Sequential 모델도 총 3개의 층으로 설정하였다. 첫 번째 층은 입력 데이터는 동일하게 독립변수의 수에 해당하는 27개를 설정하고, 은닉층(hidden layer)의 노드(node) 3개를 만들고, 활성화 함수는 동일하게 relu로 설정하였다. 두 번째 층은 은닉층(hidden layer)의 노드를 7개 만들고, 활성화 함수는 동일하게 relu로 설정하였다. 마지막 층은 출력층(output layer)의 노드는 1개로 설정하고, 마찬가지로 선형(linear) 함수를 활성화 함수로 설정하였다. 이렇게 만들어진 모델은 summary 함수로 DNN을 이용한 2017년 Sequential 모델 내부의 층 리스트를 살펴볼 수 있다. 그 형태를 표로 나타내면 〈표 4-16〉과 같다.

〈표 4-16〉 DNN을 이용한 2017년 Sequential 모델의 내부 층 리스트

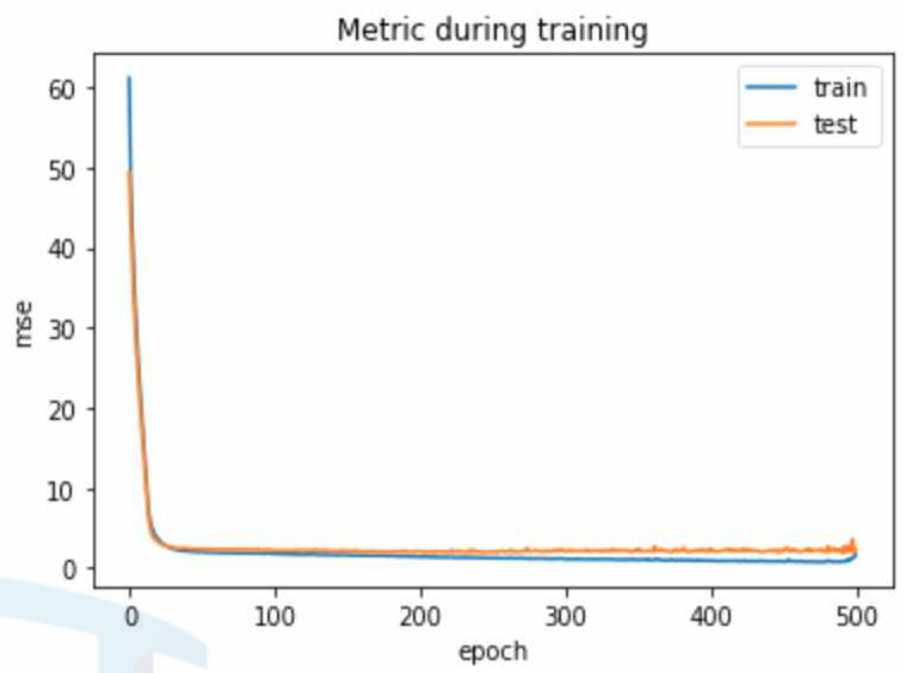
Layer (type)	Output shape	Param #
dense (Dense)	(None, 3)	84
dense_1 (Dense)	(None, 7)	28
dense_2 (Dense)	(None, 1)	8

Total params: 120
Trainable params: 120
Non-trainable params: 0

이후 2016년 모델 설계와 동일하게 2017년 공기업 재무건전성 예측모델 설계에서도 컴파일(compile) 함수를 통한 환경을 설정하고, fit 함수를 이용하여 모델을 학습시켰다.

모델 학습 결과, 2016년 모델링 결과와 마찬가지로 학습에 대한 유효한 값을 찾지 못하였다. 이 또한 데이터 수가 작음으로 인해 나타난 과적합(over-fitting) 현상으로 보인다.

DNN을 이용한 2017년 공기업 재무건전성 예측모델의 epoch, 즉 학습 프로세스가 모든 샘플에 대해 한 번 실행되는 회수에 따른 MSE(Mean Squared Error, 평균제곱오차)의 변화 그래프는 [그림 4-16]과 같다.



[그림 4-16] DNN을 이용한 2017년 예측모델의 epoch에 따른 MSE

[그림 4-1]에서 보는 것처럼 훈련 데이터셋(train dataset)도 테스트 데이터셋(test dataset)도 epoch가 얼마 되지 않아 MSE 값이 급격히 낮아졌다. 훈련 데이터셋(train dataset)과 테스트 데이터셋(test dataset) 모두 epoch 모두 MSE(Mean Squared Error, 평균제곱오차) 값이 0에 가까운 듯 하나, 훈련 데이터셋이 0에 가까이 수렴하는 것 자체가 과적합(over-fitting) 되었다고 본다. 즉, 훈련 데이터의 학습시 정확도가 거의 완벽에 가깝게 표현되지만, 실제 데이터에서는 정확도가 떨어진다.

모델 설계 결과, LightGBM과 마찬가지로 현재 본 연구의 데이터 규모로는 DNN을 이용한 공기업 재무건전성 예측 모델을 설계하는 데 적합하지 않은 것으로 나타났다.

4.4 예측모델 설계 결과 비교

본 연구에서 Random Forest, XGboost, LightGBM, DNN 4가지 기법을 이용한 예측모델 설계 결과는 다음과 같다. 4가지 기법 중 Random Forest, XGboost 두 기법을 이용하여 2016년과 2017년 공기업 재무건전성 예측모델을 설계하였고, 도출된 예측값과 실제값과 비교하여 검증할 수 있었다. 반면 LightGBM과 DNN 두 기법은 공기업 재무건전성 예측모델 설계하는데 적합하지 않은 것으로 나타났다. 이는 데이터 수가 작음으로 인해 나타난 과적합(over-fitting) 현상으로 보인다.

먼저 Random Forest, XGboost기법을 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 성능 평가를 종합하면 〈표 4-17〉과 같다.

〈표 4-17〉 Random Forest와 XGboost를 이용한 예측모델의 예측 오차

구분	2016년 공기업 재무건전성 예측모델				2017년 공기업 재무건전성 예측모델			
	MAE	MSE	RMSE	MAPE	MAE	MSE	RMSE	MAPE
Random Forest	0.839	0.980	0.989	13.05%	0.932	1.592	1.261	15.79%
XGboost	0.800	0.858	0.926	12.54%	0.800	1.121	1.058	13.14%

*MAE : Mean Absolute Error, 평균절대오차
*MSE : Mean Squared Error, 평균제곱오차
*RMSE : Root Mean Squared Error, 평균제곱근오차
*MAPE : Mean Absolute Percentage Error, 평균절대비율오차

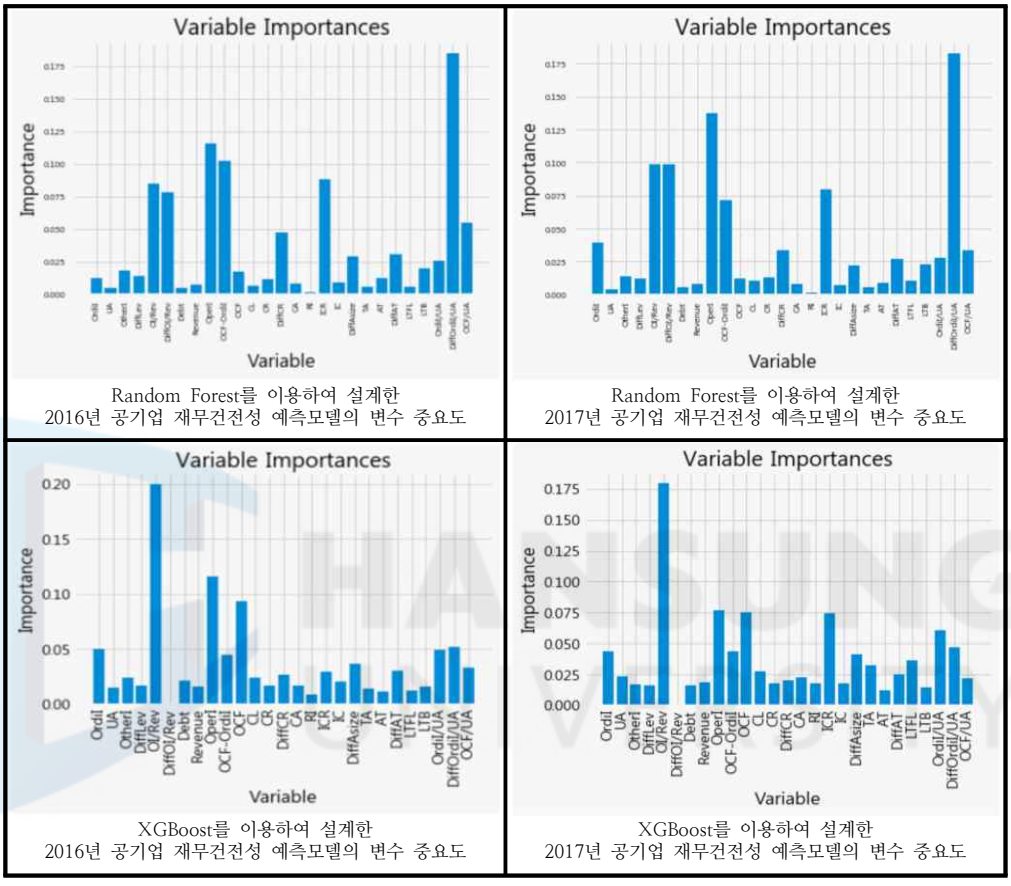
〈표 4-17〉에서 연도별 예측 모델 성능을 살펴보면, 를 이용하여 설계한 공기업 재무건전성 예측모델 모두 2016년에 비해 2017년의 예측 오차값이 높아진 것으로 나타났다. 오차 값이 크다는 의미는 예측 정확도가 떨어진다는 의미이다. 따라서 Random Forest와 XGBoost를 이용한 예측 모델 모두 2016년 비해 2017년은 예측 정확도가 떨어지는 것으로 나타났다. 이와 같은 결과는 앞서 설명한 바와 같이 2017년 정권 교체의 영향으로 보인다. 2017년 초 탄핵소추로 인한 박근혜 정부가 막을 내리고, 2017년 5월 문재인 정부가 출범하면서 정책 등 사회 전반에 많은 변화가 있었다. 특히 2017년 새정부의

탈원전 정책 발표의 영향으로 한국전력공사 외 에너지 관련 공기업 상당수가 KJY Score가 2016년 대비 2017년에 2-3점 떨어진 것으로 나타났다. KJY Score 총점이 11점임을 감안할 때 2-3점은 큰 폭이다. 또한 26개 공기업 KJY Score 평균 점수 또한 2016년 7.31점에서 2017년 6.62점으로 약 0.7점이 떨어지면서, 2012년 평균 7점대를 유지하던 KJY Score가 2017년 처음으로 평균 6점대로 떨어졌다. 이러한 변화는 기존 데이터를 통한 예측 정확도를 떨어뜨리는 요인이 될 수 있다.

모델별 예측 성능을 살펴보면, XGBoost를 이용하여 설계한 공기업 재무건전성 예측모델의 성능이 Random Forest를 이용하여 설계한 예측모델보다 더 좋은 것으로 나타났다. 즉, 예측 정확도가 더 높은 것으로 나타났다. 〈표 4-17〉에서 MAPE(Mean Absolute Percentage Error, 평균절대비율오차) 값을 살펴보면, 2016년 Random Forest 모델의 MAPE 값은 13.05%인데, XGBoost 모델의 MAPE 값은 12.54%로 0.51% 더 낮았다. 2017년 또한 Random Forest 모델의 MAPE 값은 15.79%인데, XGBoost 모델의 MAPE 값은 13.14%로 2.6% 더 낮았다. 즉, XGBoost를 이용하여 설계한 공기업 재무건전성 예측모델의 예측 정확도가 Random Forest를 이용하여 설계한 예측모델의 예측 정확도보다 더 높다고 할 수 있다.

이러한 결과는 Random Forest와 XGBoost의 트리를 구성하고 분할하는 방식에 차이로 인함이다. Random Forest는 배깅(Bagging), 임의 복원 추출인 부트스트랩(Bootstrap) 방식으로 표본을 다수 생성하여 각 결정 트리 모델에 적용하여 그 결과를 종합하는 형태이다. 임의성을 최대로 하여 결정 트리 간 상관관계를 낮추어 예측 오차가 줄이는 방식이다(하지은, 2017; 이형탁, 2019). 이에 반해 XGBoost는 이전 트리에서 얻은 정보로 다음 트리를 생성하는데 활용하여 약한 개체에 가중치를 부여하여 강한 개체로 변환시키면서 오차를 교정해 나간다. 또한 파라미터로 지정한 max_depth까지 진행한 후 손실함수(loss function)에서 개선이 일정 수준에 못 미칠 경우 역방향으로 가지치기 과정을 진행하여 트리 간 모델을 상호보완하여 오차를 줄여 트리 기반 모델에서 발생할 수 있는 과적합 문제를 해결한다(하지은, 2017). 이러한 이유로 XGBoost가 Random Forest보다 예측 정확도가 더 높은 것으로 알려져 있다.

Random Forest와 XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델의 변수 중요도 그래프를 같이 보면 [그림 4-17]과 같다.



Random Forest를 이용하여 설계한
2016년 공기업 재무건전성 예측모델의 변수 중요도

Random Forest를 이용하여 설계한
2017년 공기업 재무건전성 예측모델의 변수 중요도

XGBoost를 이용하여 설계한
2016년 공기업 재무건전성 예측모델의 변수 중요도

XGBoost를 이용하여 설계한
2017년 공기업 재무건전성 예측모델의 변수 중요도

[그림 4-17] Random forest와 XGBoost를 이용한 공기업 재무건전성 예측모델 변수 중요도 그래프

[그림 4-17]에서 보는 것처럼, 모델별 또는 연도별 변수의 중요도가 어떤 부분에서 차이가 있는지 눈으로 확인할 수 있다. 2016년과 2017년 연도별 변수 중요도가 차이가 있지만, Random Forest 설계 모델은 우측 총자산경상이익률의변화(DiffOrdiI/UA)의 변수 중요도가 가장 높은 변수임을 알 수 있으며, XGBoost의 경우 좌측 매출액영업이익률(OI/Rev)이 중요도가 가장 높은 변수임을 알 수 있다.

이 중에서 Random Forest와 XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델에서 도출된 변수 중요도 상위 4개 변수를 정리하면 〈표 4-18〉과 같다.

〈표 4-18〉 Random Forest와 XGBoost를 이용한 예측모델의 변수 중요도 상위 4개

	2016년 공기업 재무건전성 예측모델	2017년 공기업 재무건전성 예측모델
Random Forest	1. 총자산경상이익률의변화(DiffOrdiI/UA) 2. 매출액영업이익률(OI/Rev) 3 .매출영업이익률변화(DiffOI/Rev) 4. 영업이익손실(OperI)	1. 총자산경상이익률의변화(DiffOrdiI/UA) 2. 영업이익손실(OperI) 3. 매출액영업이익률(OI/Rev) 4. 매출영업이익률변화(DiffOI/Rev)
XGBoost	1. 매출액영업이익률(OI/Rev) 2. 영업이익손실(OperI) 3. 영업활동현금흐름(OCF) 4. 총자산경상이익률의변화(DiffOrdiI/UA)	1. 매출액영업이익률(OI/Rev) 2. 영업이익손실(OperI) 3. 영업활동현금흐름(OCF) 4. 이자보상배율 (ICR)

〈표 4-18〉에서 보는 바와 같이 Random Forest를 이용하여 설계한 2016년과 2017년 예측모델은 변수 중요도 일부 순위에 차이가 있으나 상위 4개 변수가 같았다. XGBoost를 이용하여 설계한 2016년과 2017년 예측모델의 경우 변수 중요도 상위 3개가 같고, 나머지 1개 변수가 다르게 나타났다. 일반적으로 예측모델에서 변수 중요도 상위 변수는 목표변수를 예측할 수 있는 주요 변수가 된다. 본 연구에서도 중요도 값이 높은 변수는 공기업 재무건전성을 예측할 수 있는 주요 변수가 된다고 볼 수 있다.

변수 중요도 상위 4개와 전체 변수를 가지고 Random Foreset와 XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델 성능 평가 결과는 〈표 4-19〉와 같다. 예측 모델의 정확도는 100-MAPE 값으로 측정하였다.

〈표 4-19〉 Random Forest와 XGBoost를 이용한 예측모델 정확도 (변수 4개 VS 변수 전체)

	2016년 공기업 재무건전성 예측모델		2017년 공기업 재무건전성 예측모델	
	변수 개수	정확도(%)	변수 개수	정확도(%)
Random Forest	변수 중요도 상위 4개	87.65	변수 중요도 상위 4개	81.90
	전체 27개	86.95	전체 27개	84.21
XGBoost	변수 중요도 상위 4개	84.50	변수 중요도 상위 4개	78.11
	전체 27개	87.46	전체 27개	86.85

*정확도(%) : 100-MAPE, MAPE(Mean Absolute Percentage Error, 평균절대비율오차)

〈표 4-19〉에서 보는 것처럼, 중요도 상위 4개 변수를 가지고 Random Forest와 XGboost로 설계한 공기업 재무건전성 모델의 예측 정확도는 전체 변수를 가지고 설계한 모델의 예측 정확도와 차이를 보였다. 4개 변수만 가지고 Random Forest를 이용한 2016년 공기업 재무건전성 예측모델만 전체 변수를 가지고 설계한 예측모델보다 예측 정확도가 더 높은 것으로 나타났다. 즉, 4개 변수로 구성된 예측모델의 예측 정확도는 87.65%였고, 전체 변수로 구성된 예측모델의 예측 정확도 86.95%였다. 이러한 결과는 상위 4개 변수가 2016년 공기업 재무건전성을 예측하는 데 유용한 지표가 됨을 의미한다.

반면 나머지는 3개의 모델은 4개 변수로 설계한 모델의 예측 정확도가 전체 변수를 가지고 설계한 모델보다 예측 정확도보다 다 낮았다. 이러한 결과는 변수 중요도를 설명하는 변수라 하더라도 상위 4개 변수만으로 공기업 재무건전성을 예측한다고 했을 때는 전체 변수로 예측했을 때보다 예측 정확도가 떨어질 수도 있음을 알 필요가 있다. 이런 경우 중요도가 높은 변수들 중 몇 개의 변수를 선택했을 때 예측의 정확도가 가장 높은지 찾아볼 필요가 있다.

V. 결 론

5.1 연구결과의 요약

본 연구에서는 기업에서도 머신러닝에 대한 관심이 급증하는 시점에서 4가지 머신러닝 기법(Random Forest, XGBoost, LightGBM, DNN)을 이용한 공기업 재무건전성 예측모델 실증연구를 통해 컨설팅 현장에서 머신러닝 기법을 적용할 수 있는 방안들을 모색해보고자 하였다.

본 연구에서 공기업 재무건전성 예측모델 실증분석 결과는 다음과 같다.

먼저 예측 모델 설계 결과, 4가지 머신러닝 기법 중 Random Forest와 XGboost 2가지 기법을 이용하여 공기업 재무건전성 예측모델을 만들고, 각 예측모델을 통해 도출된 예측값과 실제값을 비교하여 검증할 수 있었다.

하지만 나머지 2가지 기법인 LightGBM과 DNN을 통해서는 예측모델을 만들지 못하였다. 모델 학습 결과, 본 연구의 데이터셋 규모가 너무 작아 학습 결과가 유효하지 않은 것으로 확인되었다. 이는 학습 데이터를 너무 과하게 학습하는 과적합 현상 때문인 것으로 나타났다. 즉, 데이터의 수가 적다보니 기계가 학습 데이터를 거의 다 외워버려 학습 데이터에 대해서는 오차가 줄어 예측 정확도가 높아지지만, 정작 실제 데이터에 대해서는 오차가 커져 예측 정확도는 떨어지게 된다.

이와 같은 결과는 머신러닝 기법을 컨설팅 현장에 적용한다고 했을 때, 머신러닝 기법을 사용하기 위해 필요한 데이터의 양은 얼마 만큼이어야 하는지에 대한 의문이 제기될 수 있다. 일반적으로 설문 데이터를 통한 통계분석 시 모수 추정에 대한 유효한 결과를 얻기 위해서는 표본의 5배 혹은 10배의 설문이 필요하다. 머신러닝 기법의 경우, 빅데이터의 개념과 함께 등장하여 빅데이터 분석을 위한 기법으로 많이 알려져 있다. 머신러닝은 이미 1950년대 등장하였지만, 여러 번의 침체기를 겪다가 최근 몇 년 사이 빅데이터와 컴퓨터 성능의 향상으로 빠른 속도로 발전하면서 주목받기 시작했다. 특히 인터넷에 연결된 모든 기기를 통해 막대한 양의 데이터가 확보되면서 빅데이터를 기반으로 한 머신러닝의 활용 가능성은 더 높아졌다.

머신러닝의 경우, 문제의 복잡도과 머신러닝 기법의 학습 알고리즘에 따라 필요한 데이터의 양이 다르다. 즉, 원하는 수준의 모델에 필요한 데이터의 양이 표준화되어 있지 않다. 만약 예측 모델 설계 시 비슷한 프로젝트가 있는 경우 사전 연구를 통해 데이터의 수를 예측해볼 수 있지만, 그렇지 않으면 반복적인 경험을 통해 필요한 데이터의 수를 찾아가야 한다. 본 연구에서도 데이터셋의 규모가 작아 과연 머신러닝 기법을 적용한 가능할 지에 대한 의문이 있었지만, 트리 기반의 Random Forest와 XGBoost를 이용한 공기업 재무건전성 예측모델을 만들 수 있었다.

전통적인 통계분석의 경우 주로 표본을 가지고 모수를 추정하는 형태이기 때문에 적은 수의 데이터를 통한 분석이 주를 이룬다. 머신러닝 기법에서도 정답이 명시되어 있는 지도학습의 경우, 적은 수의 데이터를 통해서도 예측모델을 만들 수 있다. 반면 비지도학습의 경우, 명시적인 정답이 없이 주어진 데이터만을 가지고 규칙과 패턴을 찾는 형태이기 때문에 좀 더 많은 데이터를 필요로 한다.

한편 실제 현장에서 접하는 데이터는 거의 대부분 비선형 관계를 띄는 경우가 많다. 전통적인 통계분석에서 인과관계 추론 형태의 회귀분석 시 데이터의 선형성을 전제로 한다. 때문에 선형성을 만족시키기 위한 여러 가지 조건들을 충족해야 비로서 회귀분석을 할 수 있다. 반면 머신러닝 기법을 이용하면 선형성을 가정하지 않아도 예측모델을 만들 수 있다. 단 예측 모델을 만들기 위해서는 훨씬 더 많은 데이터를 필요로 한다. LightGBM의 경우도 기본적으로 1만개 이상의 데이터를 필요로 하며, 딥러닝은 100만개 이상의 학습데이터가 필요하다고 본다. 하지만 현재 머신러닝 기술의 발달로 DNN도 적은 수의 데이터로도 예측모델을 만들 수 있다고 알려져 있다.

본 연구에서는 트리 기반 머신러닝 기법인 Random Forest와 XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델을 만들었으며, 각각 예측정확도와 변수 중요도를 중심으로 결과를 제시하면 다음과 같다.

먼저 예측모델의 성능 평가는 모델별 예측 정확도와 연도별 예측정확도로 나누어 살펴보았다. 회귀 예측모델에서는 주로 오차 함수 값으로 예측 모델의 성능을 평가하는데, 오차 값이 낮을수록 예측 정확도가 더 높다고 본다.

모델별 예측 정확도를 살펴보면, 두 모델 다 거의 약 85% 이상의 예측 정확도를 보였다. 하지만 XGBoost가 Random Forest보다 오차 값이 낮아 예측 정확도가 더 높은 것으로 나타났다. 이는 트리 분할 방식의 차이로 인함이다. XGBoost는 약한 변수에 가중치를 부여하거나 역방향 가치지기 과정을 진행함으로 트리 간 모델을 상호보완하여 오차를 줄여 Random Forest에서 발생할 수 있는 과적합(over-fitting) 문제를 해결하기 때문이다.

인공지능 기술도 계속 발전하여 일반적으로 최신 머신러닝 기법이 이전에 나온 기법보다 예측에 있어서 더 좋은 결과를 나타낸다. 하지만 최신 기법의 적용이 무조건 옳은 것만은 아니다. 컨설팅 현장에서도 기업의 규모나 기업이 원하는 문제의 형태에 따라 적용하는 방법론이 다르듯 기업이 가지고 있는 데이터의 규모나 형태에 따라 적용 가능한 머신러닝 기법을 사용해야 한다.

한편 본 연구를 통하여 머신러닝 기법을 이용하여 설계한 모델은 동일한 형태의 새로운 데이터가 주어지면 이미 설계한 코드를 활용하여 쉽게 예측 결과를 확인할 수 있다는 장점을 확인하였다. 즉, 2016년 공기업 재무건전성 예측모델을 만든 후, 2017년 예측모델은 2016년 예측모델을 설계한 코드에 해당 연도에 해당하는 코드명만 바꾸어 입력하여 손쉽게 2017년 예측모델 결과를 바로 확인할 수 있었다.

연도별 예측 정확도를 살펴보면, XGboost와 Random Forest를 이용한 공기업 재무건전성 예측모델 모두 2016년보다 2017년도의 예측 정확도가 더 낮게 나타났다. 이러한 결과는 2017년에 있었던 정권 교체 영향으로 보인다. 2017년 초 탄핵소추로 인한 박근혜 정부가 막을 내리고, 2017년 5월 문재인 정부가 출범하면서 정책 등 사회 전반에 많은 변화 있었다.

실제 2017년 새정부의 탈원전 정책 발표의 영향으로 한국전력공사 외 에너지 관련 공기업 상당수가 재무건전성지수인 KJY Score가 2-3점 떨어진 것으로 나타났다. KJY Score의 총점이 11점임을 감안할 때 2-3점은 큰 폭이다. 또한 26개 공기업 KJY Score 평균 점수 또한 2016년 7.31점에서 2017년 6.62점으로 약 0.7점이 떨어지면서, 2012년 평균 7점대를 유지하던 KJY Score가 2017년 처음으로 평균 6점대로 떨어졌다.

이러한 변화 요인은 기존 데이터를 통한 예측 정확도를 떨어뜨리는 요인이 될 수 있다. 따라서, 예측모델 설계 시 시간에 따라 예측 시간에 따라 예측 대상의 특성이 변하는 것을 모니터링하여 그 변화를 반영하거나 주기적으로 재학습하는 작업이 필요하다(Žliobaitė, 2010) 본 연구에서는 예측 모델 설계 시 미처 그 변화 요인들을 고려하지 못하였다.

머신러닝 기법은 변수 중요도를 제공하는데, 이 변수 중요도는 상대적 중요도로 중요도 점수가 높은 독립변수일수록 목표변수 예측에 대한 기여도가 높다고 본다. 이에 Random Forest와 XGBoost를 이용한 2016년과 2017년 공기업 재무건전성 예측모델을 통해 각각 변수 중요도를 도출하였고, 각 모델별 중요도 상위 4개 변수를 뽑아 다시 예측 모델을 만들었다.

상위 4개 변수를 뽑아 만든 4개의 예측 모델 중에서 Random Forest를 이용한 2016년 공기업 재무건전성 모델의 경우, 상위 4개 변수만을 가지고 만든 예측모델이 전체변수를 가지고 만든 예측모델보다 목표변수에 대한 예측력이 높았다. 이는 중요도 상위 4개 변수만으로도 2016년 공기업 재무건전성 예측이 가능함을 의미한다. 반면 다른 3개 모델의 경우, 중요도 상위 1개 변수에서부터 10개 변수를 가지고 조합하여 예측모델을 설계해 보았으나 전체 변수를 통한 목표변수 예측력에 미치지 못하였다.

일반적으로 머신러닝 기법을 이용한 예측모델 설계 시 독립변수 전체를 사용하지 않고, 변수의 중요도를 통해 중요도가 높은 변수를 선택하여 예측모델을 설계한다. 실제 머신러닝 기법 사용 시의 독립변수(혹은 설명변수)의 개수는 경우에 따라 수 백에서 수 백만개에 이른다. 하지만 본 연구의 경우 독립변수의 수가 27개로 상대적으로 적은 수에 해당한다. 이에 전체 변수를 통해 예측모델을 만들고, 이후 변수 중요도에서 중요도가 높은 몇 개의 변수를 선택하여 만든 예측모델에서의 예측력이 어떠한지를 살펴보았다.

머신러닝 기법에서 제공하는 이 변수 중요도는 컨설팅 현장에서 유용한 도구가 될 수 있으리라 여겨진다. 한편 머신러닝 기법을 이용한 예측 모델에서 독립변수와 목표변수(혹은 종속변수) 사이의 상관관계 여부를 학습하는데, 독립변수와 목표변수의 상관관계가 반드시 인과관계를 의미하는 것은 아니다.

즉, 중요도 점수가 높은 독립변수가 목표변수에 대해 예측 정확도가 높다고 해서 이것이 두 변수 사이의 인과관계 설명해주는 것은 아니다.

연도별 공기업 재무건전성 예측모델에서 중요도가 높은 변수는 총자산경상이익률의변화, 매출액영업이익률, 매출영업이익률의변화, 영업이익손실, 영업활동현금흐름이었으며, 이 변수들은 주로 수익성과 관련된 지표들이다.

현재 정부가 시행하는 공공기관 평가에서 정성적 평가 비중이 갈수록 높아지고 있다. 기획재정부에서 발간한 공공기관 경영평가편람에 따르면, 비계량 지표 점수가 100점 만점에 2017년 45점, 2018년 52점에 이어 2019년 56점으로 매년 상승하고 있다. 숫자로 확인 가능한 재무건전성보다 정성적으로 평가하게되는 사회적가치 구현에 더 많은 배점을 두고 있기 때문이다. 이와 관련 한국전력공사가 2018년 6년만에 처음으로 1조원대 경영적자를 냈음에도 2019년 경영실적 평가에서는 양호등급인 B를 받았다. 이는 현 정부의 전면 개편된 공공기관 경영실적 평가체계에 따른 것이다. 공기업의 적자 비중이 계속해서 증가하는 상황에서 2017년부터 정성적 평가비중은 계속 늘어나고 있어 공기업 경영이 이전보다 더 방만해질 수 있다는 우려도 제기되고 있다.

이는 공기업 특성상 공익성을 추구해야하지만 공기업의 재무건전성을 높이기 위해서는 수익성을 고려해야 함을 의미한다.

5.2 연구의 시사점

본 연구의 결과를 토대로 시사점을 크게 세 가지로 정리하면 다음과 같다.

첫째, 본 연구는 전통적인 통계분석 기법이 아닌 머신러닝 기법을 이용한 예측모델 실증연구를 통해 컨설팅 현장에서 머신러닝 기법을 적용할 수 있는 방안을 모색해보았다는 점에서 의의가 있다.

불과 2-3년 사이 머신러닝에 대한 대중적 관심은 폭발적으로 증가하고 있다. 그 동안 머신러닝 기법을 활용한 연구는 주로 통계학, 의학, 공학, 자연과학 분야에서 이루어져왔으나, 최근 들어 사회과학 분야에서도 활용되기 시작했다(최필선, 민인식, 2018).

현재 글로벌 선도기업인 구글, 아마존, MS는 모든 비즈니스 역량을 데이터 및 인공지능으로 전환하고 있다. 우리나라에서도 공공기관이나 일반 기업에서도 인공지능의 활용에 대한 관심과 수요가 급증하고 있다. 또한 정부의 데이터·AI경제 활성화 계획(2019)에 따르면, 공공·민간 분야별로 데이터의 수집 및 제공을 위한 빅데이터 플랫폼과 센터를 구축할 예정이며, 중소·벤처기업이 데이터를 활용한 새로운 서비스를 개발할 수 있도록 데이터 구매 및 가공비용을 지원하는 등 다양한 사업을 추진 중이다. 이런 시점에서 본 연구는 컨설턴트들에게 머신러닝 기법의 필요와 중요성을 인식하고 활용하기 위한 시작점이 될 수 있으리라고 본다.

둘째, 본 연구를 통해 전통적 통계분석과 비교하여 머신러닝 기법의 특징과 함께 컨설턴트들이 머신러닝 기법을 실제 컨설팅 현장에서 적용 시 고려해야 할 부분들을 제시하였다는 데 의의가 있다.

현재 우리가 사는 시대를 일컬어 인공지능의 시대라고 부른다. 머신러닝은 인공지능의 한 분류로, 인공지능은 인간의 학습 능력을 머신러닝 기법을 이용하여 구현한 형태이다. 이제 인공지능은 단순 신기술이 아닌 4차 산업혁명을 촉발하는 핵심동력으로 산업구조의 변화는 물론 사회제도의 변화를 불러올 것이며, 인공지능 역량이 기업의 성장 요인이 될 것이라고 전망하고 있다(한국IR협의회, 2019).

컨설턴트는 기업의 문제해결을 위한 솔루션을 제공하는 전문가로 산업 및 경영환경의 변화에 누구보다 민감하게 반응하고 민첩하게 대응해야 한다. 이에 컨설턴트는인공지능에 대한 빠른 이해와 함께 컨설팅 현장에서의 적용 방안을 준비해야 한다. 그런 점에서 본 연구는 컨설턴트들에게 기업의 문제 해결을 위한 머신러닝 기법 적용 시 어떻게 접근해야 할지에 대한 시사점을 제공한다.

마지막으로 본 연구에서는 머신러닝 기법이 제공하는 변수의 상대적 중요도를 통해 공기업 재무건전성을 예측할 있는 주요 변수를 도출하여 공기업 재무건전성 제고에 대한 시사점을 제공하였다.

공기업 재무건전성 예측모델에서 중요도가 높은 변수는 총자산경상이익률의변화, 매출액영업이익률, 매출영업이익률의변화, 영업이익손실, 영업활동현금흐름인데, 이 변수들은 주로 수익성과 관련된 지표들이다.

이러한 결과는 최근 우리나라 공기업 부채가 계속 늘어가고 있는 상황에서 공기업의 특성상 공익성이 우선되지만, 수익성을 간과할 수 없음을 시사한다. 더불어 공기업 경영실적 평가지표 내 재무관리 지표가 점수가 줄고, 주로 정성적으로 평가하는 사회적 공헌 지표 점수 늘어나고 있는데, 공기업 경영실적 평가지표에 대한 냉정한 재고가 필요한 시점이라고 여겨진다.

5.3 연구의 한계점 및 향후 연구 방향

본 논문은 현재 기업에서도 머신러닝에 관심이 증가하고 있는 시점에서 4가지 머신러닝 기법(Random Forest, XGBoost, LightGBM, DNN)을 이용한 공기업 재무건전성 예측모델 실증연구를 통해 컨설팅 현장에서 머신러닝 기법을 적용할 수 있는 방안들을 모색해보고자 하였다. 하지만, 여러 가지 측면에서 한계점을 가지고 있다. 본 연구결과에 대한 한계점 및 향후 연구방향에 대한 제언의 내용은 다음과 같다.

첫째, 본 연구에서는 데이터의 수가 적어 4가지 머신러닝 기법 중 LightGBM이나 DNN을 통한 예측모델을 만들지 못하였다. 현재 데이터의 규모가 작은 경우에도 적용할 수 있는 기법들이 있다. 본 연구에서는 다루지 못하였으나, 추후 전이학습(Transfer Learning)이나 LSTM (Long Short-Term Memory) 기법을 이용한 예측모델 설계해보기를 권한다. 전이 학습은 데이터가 풍부한 분야에서 훈련된 모델을 다른 비슷한 분야에서 적용하는 머신러닝 기법으로, 학습 데이터가 부족한 분야 모델 구축에도 적용이 가능한 것으로 알려져 있다. 또한 LSTM은 RNN(Recurrent Neural Network, 순환신경망)의 변형된 형태이다. RNN은 연속적인 속성을 띄고 있는 번역, 문자, 음성 인식 등에서 활용되는 인공신경망이다. LSTM 기법은 직전 데이터만 고려하던 RNN의 단점을 보완하여 좀 더 거시적으로 과거 데이터를 고려하여 미래의 데이터를 예측한다. LSTM은 Hochreiter와 Schmidhber(1997)이 처음 제안하였고, 이후 많은 개선을 통해 대중화되면서 많은 분야에서 활용되고 있다.

둘째, 본 연구에서는 기획재정부를 통해 7개년 연속 선정된 27개 공기업을 전체로 하여 예측모델을 설계하였다. 하지만 기업별 특성에 따라 목표변수에 미치는 요인이 다를 수 있다. 따라서 추후 군집분석을 이용하여 공기업을 그룹별 특성으로 나누어 예측모델을 설계하여 결과를 확인해 볼 필요가 있다.

셋째, 본 연구에서 설계한 공기업 재무건전성 예측모델 모두 2016년보다 2017년도의 예측 정확도가 더 낮았는데, 이는 정권의 교체 등 외부적인 영향에 의한 것으로 파악되었다. 이러한 변화 요인은 기존 데이터를 통한 예측 정확도를 떨어뜨리는 요인이 될 수 있는데, 본 연구에서는 예측 모델 설계 시

미처 그 변화 요인들을 고려하지 못하였다. 따라서, 추후 예측모델 설계 시 예측 시간에 따라 예측 대상의 특성이 변하는 것을 모니터링하여 그 변화를 반영하거나 주기적으로 재학습하는 작업이 필요가 있다.

넷째, 본 연구에서는 공기업 재무건전성을 예측할 수 있는 주요 변수를 도출하였으나 공기업의 특성상 수익성 개선 제고 등은 정부 정책을 통해 반영되는 부분이 많아 실제 컨설턴트를 통한 현장 적용 가능성에는 한계가 있을 것으로 여겨진다. 향후 일반 기업의 데이터를 가지고 재무건전성 예측모델 설계를 통해 실제 컨설팅 현장에서 기업의 재무건전성을 제고를 위한 실제적인 적용 방안들을 찾아보고, 실제 적용 결과에 대한 연구들이 이어지기를 바란다.

다섯째, 현재 머신러닝에 대한 표준화된 방법론이 제대로 제시되어 있지 않다. 본 연구에서도 4가지 머신러닝 기법을 사용하여 예측모델을 설계하였으나, 파라미터 설정 등 표준화된 방법론이 존재하지 않아 다양한 방법으로 모델을 설계해보았다. 현재 머신러닝 기법은 계속 발전하고 있지만, 아직 전통적인 통계 분석 기법처럼 표준화된 형태의 방법론은 찾아보기 어렵다. 때문에 컨설팅 현장에서 실제 머신러닝 기법을 적용함에 있어서 다양한 시행착오를 통한 노하우를 쌓아가야 할 것이다.

여섯째, 머신러닝은 데이터를 기반으로 한다. 따라서 데이터의 양이 무조건 많다고 해서 머신러닝을 이용한 예측이 무조건 좋은 결과를 얻을 수 있는 것은 아니다. 데이터의 양보다 더 중요한 것은 필요한 데이터를 얼마나 많이 가지고 있는가 하는 점이다. 이는 컨설턴트들이 기업에서 머신러닝 기법을 적용한다고 했을 때, 기업이 목적에 맞는 정보를 제대로 수집하고 있는지를 살펴보고, 이에 대한 적절한 가이드를 해 줄 필요가 있다.

마지막으로, 현재 머신러닝에 대한 적용은 다양한 학문분야에서 이루어지고 있다. 머신러닝은 인공지능과 함께 기계 및 전자, 컴퓨터, 인문사회, 디자인, 수학, 공학 등 학제 간 융합연구분야로, 실무 적용시 해당 분야 전문가와 데이터 전문가들과의 협업이 필수적으로 여겨진다.

본 연구가 컨설턴트들에게 머신러닝의 필요와 중요성에 대한 인식과 함께 컨설팅 현장에서의 머신러닝 기법의 적용을 위한 하나의 발판이 되기를 기대해 본다.

참 고 문 헌

1. 국내문헌

고덕필a. (2003). “재무건전성지수 모형을 이용한 회사채의 신용등급예측에 관한 연구”. 『회계연구』, 8(1), 151-168.

고덕필b. (2003). “주식투자수익률 예측에 대한 F_SCORE 모형의 유용성과 현실적합성”. 『회계연구』, 8(1), 1-18.

관계부처 합동. (2019). 데이터·AI경제 활성화 계획(‘19~’23년).

국가법령정보센터. (2019). 공공기관의 운영에 관한 법률, http://www.law.go.kr/법령/공공기관의운영에관한법률.

국회예산정책처a. (2019). 『2018회계연도 공공기관 결산분석 I』. 국회예산정책처 발간자료.

국회예산정책처b. (2019). 『2019 대한민국 공공기관』. 국회예산정책처 발간자료.

권안나. (2019). “랜덤포레스트를 이용한 변수 선택”. 인하대학교 대학원. 석사학위논문.

기획재정부a. (2011). 『2011년도 공공기관 경영평가편람』. 기획재정부 발간자료.

기획재정부b. (2011). 『2011회계연도 공기업·준정부기관 결산서』. 기획재정부 발간자료.

기획재정부a. (2012). 『2012년도 공공기관 경영평가편람』. 기획재정부 발간자료.

기획재정부b. (2012). 『2012회계연도 공기업·준정부기관 결산서』. 기획재정부 발간자료.

기획재정부a. (2013). 『2013년도 공공기관 경영평가편람』. 기획재정부 발간자료.

기획재정부b. (2013). 『2013회계연도 공기업·준정부기관 결산서』. 기획재정부 발간자료.

기획재정부a. (2014). 『2014년도 공공기관 경영평가편람』. 기획재정부 발간자료.

기획재정부b. (2014). 『2014회계연도 공기업·준정부기관 결산서』. 기획재정부 발간자료.

기획재정부a. (2015). 『2015년도 공공기관 경영평가편람』. 기획재정부 발간자료.

기획재정부b. (2015). 『2015회계연도 공기업·준정부기관 결산서』. 기획재정부 발간자료.

기획재정부a. (2016). 『2016년도 공공기관 경영평가편람』. 기획재정부 발간자료.

기획재정부b. (2016). 『2016회계연도 공기업·준정부기관 결산서』. 기획재정부 발간자료.
기획재정부a. (2017). 『2017년도 공공기관 경영평가편람』. 기획재정부 발간자료.
기획재정부b. (2017). 『2017회계연도 공기업·준정부기관 결산서』. 기획재정부 발간자료.
기획재정부. (2018). 『2018년도 공공기관 경영평가편람』. 기획재정부 발간자료.
기획재정부. (2019). 『2019년도 공공기관 경영평가편람』. 기획재정부 발간자료.
김동영, 원동은. (2019). 『코스피 단기예측 AI 모델 - 랜덤 포레스트 기법을 활용한 머신러닝 기반 모델』. Quantitative Issue. 서울: 삼성증권.
김성진. (2015). "랜덤 포레스트를 활용한 기업채권등급평가 모형". 국민대학교 비즈니스IT전문대학원. 석사학위논문.
김승연, 정용주. (2017). 『처음 배우는 머신러닝』. 서울: 한빛미디어.
김재휘, 김재희. (2019). "앙상블 학습과 온도 변수를 이용한 A호텔의 전력소모량 예측". 『응용통계연구』. 32(2), 319-330.
김종영. (2019). "부스팅 기반 최신 트리 알고리즘 비교 연구". 서강대학교 대학원. 석사학위논문.
김태진, 홍정식, 전윤수, 박종률, 안태욱. (2018). "랜덤포레스트를 이용한 모기업의 하향 거래처 기업의 분류: 자동차 부품산업의 가치사슬을 중심으로". 『한국전자거래학회지』. 23(1), 1-22.
김현주. (2019). "빅데이터 및 인공지능을 활용한 지방지역 교통수요예측". 서울시립대학교 국제도시과학대학원. 석사학위논문.
김형민. (2003). "재무건전성지수를 이용한 신용등급평가와 주가수익률 예측에 관한 연구", 청주대학교 대학원, 박사학위논문.
머피, K. (2015). 『머신러닝 Machine Learning』. 서울: ㈜에이콘출판.
뮐러, A. & 가이도, S. (2017). 『파이썬 라이브러리를 활용한 머신러닝』. 서울: 한빛미디어.
민성욱. (2017). "딥러닝(Deep Learning)을 이용한 주택가격 예측모형 연구". 강남대학교 대학원. 박사학위논문.

박경덕, 한길석, 윤석진. (2008). “신용등급예측을 위한 수정재무건전성지수모형 도출에 관한 연구”. 『국제회계연구』, 21(9), 1-18.

박원기. (2018). “딥러닝 소개와 금융업 적용 사례”. 고려대학교 정책대학원. 석사학위논문.

박유선. (2018). “기계학습 기반 침해사고 경보 위험도 예측”. 고려대학교 대학원. 석사학위논문.

박인근, 홍지후, 강남규, 김성호, 정구범. (2019). 『4차 산업혁명 현장 전문가가 알려주는 빅데이터 분석과 활용』. 파주: ㈜제이펍.

박진, 최진욱, 박진희, 김지영, 허경선. (2012). 『공공기관 부채의 잠재적 위험성 분석과 대응방안』. 한국조세연구원.

백유진. (2019). 『ModAugNet: 과적합 방지 LSTM 모듈과 예측 LSTM 모듈을 활용한 새로운 주가지수 예측 모델』. 아주대학교 대학원, 석사학위논문.

송현화. (2019). 『Python 데이터 분석 실무』, WikiDocs.

심재헌. (2016). 머신러닝 알고리즘을 이용한 부동산가치 산정에 관한 소고. 『부동산포커스』. 100, 52-58. 서울: 한국감정원.

알리오플러스 사이트. 기관정보. 공공기관이란?, http://alioplus.go.kr/organi-zation /organByPub.do.

양진용. (2017). “기업 재무 정보를 활용한 머신 러닝 기반 경영 예측 시스템”, 한성대학교 대학원. 박사학위논문.

원승현. (2017). “딥러닝 기법을 활용한 매립가스 발전소의 메탄가스 농도 및 전력 생산량 예측”. 서울과학기술대학교 에너지환경대학원, 박사학위논문

윤운규. (2018). “적합과 내부 공변량 변화를 줄이기 위해 제곱 가중치를 사용한 완전 연결 층 인공신경망”. 영남대학교 대학원, 석사학위논문.

이근영. (2015). “머신러닝을 활용한 스마트 서비스와 금융”. 『전자금융과 금융보안』(31-66). 서울 : 금융보안원.

이용주. (2018). “딥러닝 기반의 링크통행시간을 이용한 차량대기길이 추정모형

개발". 아주대학교 일반대학원, 박사학위논문.

이장희, 이종열. (2013). 재무건전성지수와 회계이익의 질과의 관련성 분석. 『회계연구』, 18(3): 21-47.

이현상, 오세환. (2019). 머신러닝 기법을 활용한 기업 신용 평점 예측모델 개발. 한국경영정보학회 학술대회논문집.『한국경영정보학회』. 2019(5), 293-299

이형탁. (2019). "머신러닝 예측 알고리즘을 이용한 선박 접안속도에 영향을 미치는 요인 분석". 한국해양대학교 대학원. 석사학위논문.

전해남. (2018). "융합보안과 인공지능 - 머신러닝, 딥러닝". SK infosec 공식블로그. http://blog.skinfosec.com/221396731850.

정준수. (2017). "재무건전성지수를 이용한 공공기관 재무건전성 제고방안". 한성대학교 대학원. 박사학위논문.

제롱, O. (2018). 『핸즈온 머신러닝』. 서울: 한빛미디어.

조태호. (2017). 『모두의 딥러닝』, 서울: 길벗.

최정원, 오세경, 장재원. (2017). 빅데이터와 인공지능 기법을 이용한 기업 부도예측 연구. 한국재무학회 학술대회, 17(11), 396-435.

최종고. (2019). "머신러닝을 활용한 건설근로자 중대사고 위험성 예측 모델", 성균관대학교 일반대학원. 석사학위논문.

최필선, 민인식. (2018). 머신러닝 기법을 이용한 대졸자 취업예측 모형. 『직업능력개발연구』, 21(1), 31～54.

표영호. (2012). "에너지 관련 공공기관의 국제회계기준 도입 과정과 재무제표에 미치는 효과에 대한 사례 연구". 성균관대학교 경영전문대학원. 석사학위논문.

하지은. (2017). "RandomForest와 XGBoost를 활용한 민원 카테고리 및 담당부서 자동분류 성능 비교". 연세대학교 정보대학원. 석사학위논문.

한국IR협의회. (2019). 산업테마보고서『인공지능』.

한은정. (2015). "건강검진 자료에서 Random forests를 이용한 백내장 발생 위험군 예측모형", 연세대학교 대학원. 석사학위논문.

2. 국외문헌

Breiman. L. (1996). "Bagging Predictors. *Machine Learning"*. 24(2), 123-140.

Breiman. L. (2001). "RANDOM FORESTS. *Machine Learning"*. 45(1), 5-32.

Breiman, L. & Cutler, A. (2014). 『Random Forests』. https://www.stat.berkeley.edu/~breiman/RandomForests.

Chen, T. (2014) 『Introduction to Boosted Trees』. UNIVERSITY of WATSHINTON, https://homes.cs.washington.edu/~tqchen/pdf/BoostedTree.pdf

Chen, T. & Guestrin, C. (2016) 『XGBoost: A Scalable Tree Boosting System』. In KDD'16 Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785-794, New York: ACM.

Claesen, Marc & Moor, B. D. (2015). 『Hyperparameter Search in Machine Learning』. MIC 2015: The XI Metaheuristics International Conference. Agadir: Morocco, arXiv:1502.02127 [cs.LG]

Friedman, J. H. (2001). "Greedy Function Approximation: A Gradient Boosting Machine". *The Annals of Statistics*, 29, 1189-1232.

Hebb, D. O. (1949). 『The Organization of Behavior』. New York: Wiley & Sons.

Hinton G. E., Osindero S. & Teh Y. W. (2006). "A fast learning algorithm for deep belief nets". *Neural computation*, 18(7), 1527-1554.

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W,. Ma, W., Ye, Q., Liu T. Y. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. In Advances in Neural Information Processing Systems 30 (NIPS 2017), 3149-3157.

Li. R. H. & Belford. G. G. (2002). Instability of Decision Tree Classification Algorithms. In KDD'02 Proceedings of the eighth ACM SIGKDD

international conference on Knowledge discovery and data mining. 570-575, New York: ACM.

McCulloch, W. S. & Pitts, W. (1943). "A logical calculus of the ideas immanent in nervous activity". *Bulletin of mathematical biophysics*. 5, 115~133

Minsky, M. & Papert, S. A. (1969). 『Perceptrons: An Introduction to Computational Geometry』. Cambridge MA. MIT Press

Opitz & Maclin. (1999). "Popular Ensemble Methods: An Empirical Study". *Journal of Artificial Intelligence Research*, 11, 169-198.

Piotroski, Joseph. D. (2000). "Value Investing: The Use of Historical Financial Statement Information to Separate Winners from Losers". *Journal of Accounting Research*, 38. 1-41.

Rosenblatt, F. (1958). "The perceptron: A probabilistic model foe information storage and organization in the brain". *Psychological Review*, 65(6), 386-408

Samuel, A. L. (1595). "Some Studies in Machine Learning Using the Game of Checkers". *IBM Journal of Research and Development,* 3(3), 210-229.

Siroky, D. S. (2009). "Navigating Random Forests and related advances in algorithmic modeling". *Statistics Surveys,* 3, 147-163.

Wu, C., Buyya, R. & Ramamohanarao, K. (2016). "Big Data Analytics = Machine Learning + Cloud Computing. a Book Chapter in "Big Data: Principles and Paradigms, R. Buyya, R. Calheiros, and A. Dastjerdi (eds)", 1-27. Massachusetts: Morgan Kaufmann.

Žliobaitė, I. (2010). "Learning under Concept Drift: an Overview". eprint arXiv:1010.4784, 1-36.

부 록

공기업 재무건전성 예측모델 설계를 위해 사용한 공기업 재무자료 세부항목 (2017년 회계연도 자료)

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (1/10)

명칭	년도	유동자산	전기유동자산	자산총계	전기자산총계
공기업1	2017	8,658,040,826,789	8,658,040,826,789	37,139,439,444,281	40,042,003,026,119
공기업2	2017	19,708,526,040,826	19,708,526,040,826	181,788,914,780,669	177,837,042,097,255
공기업3	2017	592,945,210,582	592,945,210,582	12,356,809,907,459	10,952,038,497,514
공기업4	2017	774,804,082,429	774,804,082,429	4,564,559,499,697	4,432,613,876,121
공기업5	2017	209,810,178,096	209,810,178,096	5,645,707,829,201	5,620,436,304,404
공기업6	2017	55,360,768,050	55,360,768,050	2,904,331,325,510	2,845,408,997,022
공기업7	2017	162,593,501,246	162,593,501,246	391,165,581,028	387,633,623,093
공기업8	2017	780,782,320,317	780,782,320,317	1,402,082,990,383	1,386,501,221,138
공기업9	2017	651,463,409,988	651,463,409,988	2,602,153,545,640	2,528,907,311,873
공기업10	2017	529,880,177,309	529,880,177,309	4,151,834,734,663	4,365,775,105,846
공기업11	2017	218,165,208,492	218,165,208,492	799,242,644,434	734,126,655,478
공기업12	2017	1,609,241,901,188	1,609,241,901,188	19,492,001,642,355	22,067,195,680,615
공기업13	2017	446,926,242,920	446,926,242,920	5,721,590,908,623	5,133,804,856,969
공기업14	2017	0	0	6,690,084,377,021	6,348,880,887,819
공기업15	2017	579,780,516,507	579,780,516,507	1,358,388,570,100	1,152,987,028,306
공기업16	2017	122,896,083,078	122,896,083,078	248,358,124,552	243,005,163,897
공기업17	2017	946,221,851,264	946,221,851,264	61,089,747,798,848	59,582,562,057,520
공기업18	2017	6,422,010,201,048	6,422,010,201,048	20,867,295,406,528	20,299,429,822,127
공기업19	2017	1,096,123,085,951	1,096,123,085,951	19,878,343,051,393	18,512,929,447,331
공기업20	2017	86,772,415,395,418	86,772,415,395,418	173,682,867,490,288	172,321,732,989,915
공기업21	2017	6,063,171,372,829	6,063,171,372,829	55,232,689,988,427	53,022,477,381,617
공기업22	2017	1,541,010,532,246	1,541,010,532,246	10,031,379,745,489	9,877,452,174,949
공기업23	2017	1,216,142,619,023	1,216,142,619,023	10,249,534,235,636	9,442,509,349,242
공기업24	2017	1,174,858,725,055	1,174,858,725,055	9,628,670,614,412	9,793,102,628,861
공기업25	2017	1,256,549,224,591	1,256,549,224,591	10,026,394,288,040	10,065,791,494,126
공기업26	2017	993,868,813,403	993,868,813,403	8,926,565,686,345	9,045,177,460,945

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (2/10)

명칭	전전기자산총계	유동부채	전기유동부채	장기차입금	전기장기차입금
공기업1	42,385,339,683,322	6,590,430,029,281	5,556,281,611,466	162,507,930,130	420,679,520,566
공기업2	175,257,358,220,650	23,424,281,129,279	24,739,225,543,985	2,455,736,763,868	1,799,750,259,020
공기업3	9,437,511,842,402	1,307,653,463,904	701,725,442,511	90,000,000,000	146,028,400,000
공기업4	4,336,744,531,024	245,446,430,170	201,315,564,445	0	0
공기업5	5,542,322,139,855	346,231,955,339	226,133,412,551	0	0
공기업6	2,801,627,642,953	72,478,452,451	27,592,590,030	0	0
공기업7	378,887,706,402	40,026,454,087	43,397,335,763	25,177,900,000	25,076,375,000
공기업8	1,345,327,073,180	198,433,843,725	209,565,680,539	85,694,444,389	85,500,000,000
공기업9	2,506,632,106,846	82,032,123,528	72,329,261,775	0	0
공기업10	4,687,481,355,394	879,720,688,138	575,688,755,932	782,244,079,331	615,992,576,588
공기업11	730,323,102,438	1,000,397,636,698	1,244,490,395,181	550,702,067,000	200,812,602,000
공기업12	23,205,165,061,026	4,251,081,769,506	4,432,721,060,630	1,035,264,579,008	1,096,273,083,249
공기업13	4,843,831,863,638	977,673,870,522	756,492,467,126	0	0
공기업14	5,579,794,342,614	0	0	0	0
공기업15	1,041,800,551,220	171,253,665,311	125,564,072,822	301,404,344,759	332,485,959,083
공기업16	233,369,234,639	24,481,833,790	26,067,050,283	0	0
공기업17	57,605,701,102,895	3,739,537,081,919	4,653,532,083,386	600,000,000,000	400,000,000,000
공기업18	19,550,638,906,238	3,056,094,914,289	3,154,565,108,171	344,324,609,371	408,438,471,607
공기업19	18,198,705,287,676	2,854,490,504,908	2,908,664,022,494	61,000,000,000	116,000,000,000
공기업20	169,889,480,479,777	52,901,712,998,278	51,744,728,893,467	31,743,990,400,453	31,175,248,273,323
공기업21	51,252,786,827,890	2,733,904,020,159	3,188,161,666,294	183,647,813,900	208,084,973,900
공기업22	9,387,857,507,444	1,745,450,629,292	1,529,614,548,378	171,856,198,000	147,201,834,800
공기업23	8,114,548,381,983	1,246,420,460,620	1,088,862,370,455	251,158,251,806	264,714,071,255
공기업24	9,207,872,135,261	1,297,360,718,281	1,650,500,555,179	1,548,130,000	2,059,440,000
공기업25	9,260,455,052,327	1,516,266,913,909	1,489,552,977,019	288,747,521,496	191,984,950,000
공기업26	8,880,573,960,785	1,422,373,825,129	,752,093,005,599	78,330,206,046	61,646,795,379

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (3/10)

명칭	사채	전기사채	수익매출액	전기수익매출액	영업이익손실
공기업1	18,580,960,200,000	21,152,023,000,000	22,172,305,487,411	21,108,115,739,226	1,033,936,616,381
공기업2	43,270,824,912,572	43,012,959,503,373	59,814,861,528,069	60,190,384,637,354	4,953,152,072,485
공기업3	1,710,000,000,000	1,760,000,000,000	2,499,104,360,348	2,241,343,498,014	1,464,147,417,430
공기업4	0	0	883,195,784,876	830,296,699,245	227,327,507,901
공기업5	1,420,000,000,000	1,550,000,000,000	341,417,665,835	350,270,814,962	142,489,779,412
공기업6	670,000,000,000	670,000,000,000	147,646,931,868	136,456,452,480	35,002,453,961
공기업7	0	0	484,874,566,365	469,921,936,716	13,061,508,858
공기업8	0	0	732,967,476,061	779,874,197,596	66,045,934,598
공기업9	0	0	7,844,702,102,459	7,789,790,982,992	207,745,698,525
공기업10	3,661,867,000,000	3,792,723,000,000	722,869,431,304	330,481,001,748	-178,697,900,422
공기업11	170,000,000,000	170,000,000,000	138,127,130,669	154,958,514,051	-51,381,784,077
공기업12	9,044,863,282,692	10,289,014,350,237	2,312,485,962,183	2,430,397,594,201	175,871,217,525
공기업13	2,160,000,000,000	1,920,000,000,000	1,834,422,361,734	1,719,867,257,811	119,796,039,704
공기업14	0	0			
공기업15	0	0	728,147,740,884	649,908,399,552	211,283,385,700
공기업16	0	0	142,737,635,834	141,687,604,975	11,311,725,583
공기업17	22,032,663,560,000	21,610,256,290,000	8,059,818,660,507	8,159,035,729,436	971,591,524,027
공기업18	9,937,819,000,000	9,802,014,500,000	3,375,560,292,357	3,618,084,599,554	427,716,418,374
공기업19	10,317,932,870,000	9,560,429,700,000	5,786,708,745,152	5,693,642,953,998	-469,912,245,057
공기업20	31,749,795,900,000	37,554,697,900,000	23,559,413,323,511	22,967,747,894,973	3,001,428,509,373
공기업21	7,523,510,000,000	7,872,525,000,000	9,510,942,713,097	11,277,136,267,254	1,397,225,114,351
공기업22	2,372,847,000,000	2,635,116,250,000	5,399,328,370,684	5,101,915,297,329	270,650,199,908
공기업23	4,472,465,000,000	3,975,100,000,000	4,260,671,763,481	3,817,348,716,424	
공기업24	4,110,364,000,000	3,933,066,000,000	4,222,449,292,599	4,179,781,770,903	361,484,381,350
공기업25	3,590,000,000,000	3,822,550,000,000	4,495,818,426,123	4,218,813,486,696	266,211,563,818
공기업26	2,381,400,000,000	2,378,500,000,000	4,669,824,083,859	4,247,375,220,251	422,570,095,071

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (4/10)

명칭	전기영업이익손실	영업수익	전기영업수익	영업이익	전기영업이익
공기업1	998,213,959,705				
공기업2	12,001,599,476,721				
공기업3	1,308,077,332,539				
공기업4	235,880,215,999				
공기업5	178,123,058,879				
공기업6	32,988,225,727				
공기업7	9,024,635,774				
공기업8	100,954,501,359				
공기업9	204,091,207,304				
공기업10	-313,589,053,156				
공기업11	-46,827,763,123				
공기업12	-232,303,286,723				
공기업13	161,656,016,220				
공기업14		817,194,736,394	875,499,879,733	675,229,703,542	523,804,441,901
공기업15	143,911,359,876				
공기업16	17,454,674,657				
공기업17	725,488,814,667				
공기업18	364,099,428,162				
공기업19	121,583,845,145				
공기업20	3,175,696,174,276				
공기업21	3,847,243,089,524				
공기업22	834,055,677,549				
공기업23				195,607,504,087	520,442,908,201
공기업24	588,660,923,319				
공기업25	603,371,690,230				
공기업26	672,116,947,597				

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (5/10)

명칭	기타이익손실	전기기타이익손실	이자비용	유상증자	영업활동으로인한현금흐름
공기업1	-1,699,967,754,661	-1,071,199,534,328	829,544,824,571	0	2,507,597,962,189
공기업2	156,626,577,598	70,498,260,427	1,789,552,030,253	17,002,149,854	11,249,893,861,796
공기업3	42,514,125,425	-2,469,607,415	26,678,650,311	0	1,593,522,809,158
공기업4	13,226,255,050	-488,753,828	0	0	262,376,931,750
공기업5	0	0	46,238,339,230	0	172,213,736,414
공기업6	8,762,595,682	751,704,171	20,509,466,981	0	44,073,047,925
공기업7	376,447,599	470,128,362	1,273,127,312	0	31,049,467,622
공기업8	18,184,971,412	4,481,700,199	1,720,990,314	0	63,972,500,476
공기업9	-1,634,073,148	-3,185,749,515	0	0	232,304,421,065
공기업10	-145,035,259,880	-562,579,393,904	2,557,645,233	77,250,379,073	-144,064,280,538
공기업11	832,145,609	1,691,520,240	30,548,346,360	33,720,000,000	-126,502,777,225
공기업12	-463,149,783,883	-422,718,977,974	401,018,824,688	88,013,000,000	695,382,273,517
공기업13	3,448,923,923	8,311,193,660	44,925,121,720	0	445,412,084,951
공기업14	650,530	-961,458	4,897,147,763	0	221,154,465,625
공기업15	1,517,533,117	114,989,017	13,651,784,614	0	241,832,213,422
공기업16	178,526,125	-358,116,739	0	0	6,020,177,484
공기업17	-11,034,660,359	299,023,697,290	798,367,840,990	1,360,596,000,000	1,829,983,809,690
공기업18	-1,889,366,066	-143,011,413,145	262,545,872,445	14,427,155,200	220,497,767,843
공기업19	-140,271,290,846	56,298,826,624	392,321,139,493	142,306,955,000	508,777,866,527
공기업20	-3,851,803,059	-23,809,656,227	139,035,425,449	1,351,362,935,000	11,556,590,089,939
공기업21	10,300,303,507	11,607,736,495	496,270,879,479	22,305,774	2,312,082,495,163
공기업22	18,716,240,790	-2,754,087,545	91,797,388,983	1,890,326,026	527,514,547,216
공기업23	-10,429,242,086	-4,994,943,006	76,124,048,879	0	474,128,981,597
공기업24	-9,350,718,831	4,814,609,628	122,358,908,403	0	583,115,968,891
공기업25	-10,244,060,609	-17,893,318,598	112,167,331,782	0	654,742,307,394
공기업26	1,712,124,678	-10,671,100,075	99,135,172,702	0	558,862,005,767

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (6/10)

명칭	장기금융부채	전기장기금융부채	경상이익	전기경상이익	기초자산
공기업1	18,743,468,130,130	21,572,702,520,566	-666,031,138,280	-72,985,574,623	38,590,721,235,200
공기업2	45,726,561,676,440	44,812,709,762,393	5,109,778,650,083	12,072,097,737,148	179,812,978,438,962
공기업3	1,800,000,000,000	1,906,028,400,000	1,506,661,542,855	1,305,607,725,124	11,654,424,202,487
공기업4	0	0	240,553,762,951	235,391,462,171	4,498,586,687,909
공기업5	1,420,000,000,000	1,550,000,000,000	142,489,779,412	178,123,058,879	5,633,072,066,803
공기업6	670,000,000,000	670,000,000,000	43,765,049,643	33,739,929,898	2,874,870,161,266
공기업7	25,177,900,000	25,076,375,000	13,437,956,457	9,494,764,136	389,399,602,061
공기업8	85,694,444,389	85,500,000,000	84,230,906,010	105,436,201,558	1,394,292,105,761
공기업9	0	0	206,111,625,377	200,905,457,789	2,565,530,428,757
공기업10	4,444,111,079,331	4,408,715,576,588	-323,733,160,302	-876,168,447,060	4,258,804,920,255
공기업11	720,702,067,000	370,812,602,000	-50,549,638,468	-45,136,242,883	766,684,649,956
공기업12	10,080,127,861,700	11,385,287,433,486	-287,278,566,358	-655,022,264,697	20,779,598,661,485
공기업13	2,160,000,000,000	1,920,000,000,000	123,244,963,627	169,967,209,880	5,427,697,882,796
공기업14	0	0	675,230,354,072	523,803,480,443	6,519,482,632,420
공기업15	301,404,344,759	32,485,959,083	212,800,918,817	144,026,348,893	1,255,687,799,203
공기업16	0	0	11,490,251,708	17,096,557,918	245,681,644,225
공기업17	2,632,663,560,000	22,010,256,290,000	960,556,863,668	1,024,512,511,957	60,336,154,928,184
공기업18	10,282,143,609,371	10,210,452,971,607	425,827,052,308	221,088,015,017	20,583,362,614,328
공기업19	10,378,932,870,000	9,676,429,700,000	-610,183,535,903	177,882,671,769	19,195,636,249,362
공기업20	63,493,786,300,453	68,729,946,173,323	2,997,576,706,314	3,151,886,518,049	173,002,300,240,101
공기업21	7,707,157,813,900	8,080,609,973,900	1,407,525,417,858	3,858,850,826,019	54,127,583,685,022
공기업22	2,544,703,198,000	2,782,318,084,800	289,366,440,698	831,301,590,004	9,954,415,960,219
공기업23	4,723,623,251,806	4,239,814,071,255	185,178,262,001	515,447,965,195	9,846,021,792,439
공기업24	4,111,912,130,000	3,935,125,440,000	352,133,662,519	593,475,532,947	9,710,886,621,637
공기업25	3,878,747,521,496	4,014,534,950,000	255,967,503,209	585,478,371,632	10,046,092,891,083
공기업26	2,459,730,206,046	2,440,146,795,379	424,282,219,749	661,445,847,522	8,985,871,573,645

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (7/10)

명칭	전기기초자산	총자산경상이익률	전기총자산경상이익률	총자산경상이익률의변화	총자산대비영업현금흐름비율
공기업1	41,213,671,354,721	-0.017258841	-0.001770907	-0.015487935	0.064979298
공기업2	176,547,200,158,952	0.028417185	0.068378868	-0.039961683	0.062564415
공기업3	10,194,775,169,958	0.129278076	0.128066358	0.001211719	0.136731149
공기업4	4,384,679,203,573	0.053473186	0.053684991	-0.000211805	0.058324303
공기업5	5,581,379,222,130	0.025295217	0.031913807	-0.006618590	0.030571904
공기업6	2,823,518,319,988	0.015223313	0.011949605	0.003273708	0.015330448
공기업7	383,260,664,748	0.034509425	0.024773646	0.009735779	0.079736773
공기업8	1,365,914,147,159	0.060411234	0.077190943	-0.016779710	0.045881706
공기업9	2,517,769,709,360	0.080338796	0.079795009	0.000543786	0.090548301
공기업10	4,526,628,230,620	-0.076015024	-0.193558738	0.117543713	-0.033827396
공기업11	732,224,878,958	-0.065932764	-0.061642597	-0.004290168	-0.164999752
공기업12	22,636,180,370,821	-0.013825030	-0.028936961	0.015111931	0.033464663
공기업13	4,988,818,360,304	0.022706673	0.034069633	-0.011362959	0.082062800
공기업14	5,964,337,615,217	0.103571156	0.087822574	0.015748582	0.033922088
공기업15	1,097,393,789,763	0.169469608	0.131243998	0.038225609	0.192589443
공기업16	238,187,199,268	0.046768865	0.071777820	-0.025008955	0.024503978
공기업17	58,594,131,580,208	0.015920087	0.017484900	-0.001564812	0.030329805
공기업18	19,925,034,364,183	0.020687925	0.011095992	0.009591933	0.010712427
공기업19	18,355,817,367,504	-0.031787617	0.009690806	-0.041478424	0.026504871
공기업20	171,105,606,734,846	0.017326803	0.018420709	-0.001093906	0.066800211
공기업21	52,137,632,104,754	0.026003847	0.074012775	-0.048008928	0.042715420
공기업22	9,632,654,841,197	0.029069153	0.086300361	-0.057231208	0.052993018
공기업23	8,778,528,865,613	0.018807419	0.058716896	-0.039909476	0.048154371
공기업24	9,500,487,382,061	0.036261742	0.062467904	-0.026206162	0.060047655
공기업25	9,663,123,273,227	0.025479309	0.060588937	-0.035109629	0.065173826
공기업26	8,962,875,710,865	0.047216591	0.073798396	-0.026581805	0.062193411

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (8/10)

명칭	영업현금흐름과경상이익의차이	레버리지변화	유동비율	전기유동비율	유동비율의변화
공기업1	0.082238139	-0.073313851	1.150484856	1.558243702	-0.407758846
공기업2	0.034147230	0.005082236	0.817173723	0.796650890	0.020522834
공기업3	0.007453072	-0.009097695	0.670937560	0.844981776	-0.174044216
공기업4	0.004851117	0	2.348782555	3.848704319	-1.499921764
공기업5	0.005276687	-0.023077993	0.099571438	0.927815911	-0.828244473
공기업6	0.000107135	0	0.508543611	2.006363592	-1.497819981
공기업7	0.045227348	0.000260722	4.288582045	3.746624035	0.541958010
공기업8	-0.014529528	0.000139457	4.216522030	3.725716531	0.490805499
공기업9	0.010209505	0	9.270192930	9.006913578	0.263279352
공기업10	0.042187628	0.008311135	0.529453583	0.920428221	-0.390974638
공기업11	-0.099066988	0.456366858	0.289332359	0.175304855	0.114027504
공기업12	0.047289693	-0.062809662	0.380640046	0.363037033	0.017603013
공기업13	0.059356126	0.044217642	0.632025340	0.590787433	0.041237906
공기업14	-0.069649068	0	0	0	-0.000000001
공기업15	0.023119835	-0.024752661	3.952914940	4.617407699	-0.664492758
공기업16	-0.022264888	0	5.140471926	4.714614110	0.425857815
공기업17	0.014409717	0.010315660	0.196451069	0.203334120	-0.006883051
공기업18	-0.009975498	0.003482941	2.304184333	2.035783058	0.268401276
공기업19	0.058292488	0.036597024	0.600274135	0.376847610	0.223426525
공기업20	0.049473408	-0.030266418	1.596475904	1.676932458	-0.080456554
공기업21	0.016711573	-0.006899480	2.110276486	1.901776637	0.208499849
공기업22	0.023923865	-0.023870299	0.969483068	1.007450232	-0.037967165
공기업23	0.029346951	0.049137529	1.087533429	1.116892871	-0.029359442
공기업24	0.023785913	0.018205000	0.821744056	0.711819649	0.109924407
공기업25	0.039694517	-0.013516442	0.712377911	0.843574713	-0.131196802
공기업26	0.014976821	0.002179356	0.775346979	0.567246607	0.208100372

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (9/10)

명칭	유상증자여부	매출액영업이익률	전기매출액영업이익률	매출액영업이익률의변화	자산회전율
공기업1	1	0.046631895	0.047290529	-0.000658634	0.574550171
공기업2	0	0.082808050	0.199393965	-0.116585915	0.332650413
공기업3	1	0.585868858	0.583613058	0.002255800	0.214433962
공기업4	1	0.257391976	0.284091477	-0.026699502	0.196327390
공기업5	1	0.417347412	0.508529547	-0.091182135	0.060609497
공기업6	1	0.237068617	0.241749108	-0.004680491	0.051357774
공기업7	1	0.026937913	0.019204542	0.007733371	1.245185059
공기업8	1	0.090107592	0.129449726	-0.039342134	0.525691477
공기업9	1	0.026482293	0.026199831	0.000282462	3.057731070
공기업10	0	-0.247206332	-0.948886779	0.701680447	0.169735277
공기업11	0	-0.371989079	-0.302195484	-0.069793595	0.180161597
공기업12	0	0.076052880	-0.095582421	0.171635301	0.111286363
공기업13	1	0.065304502	0.093993310	-0.028688807	0.337974294
공기업14	1	0.826277598	0.598291849	0.227985749	0.125346562
공기업15	1	0.290165544	0.221433297	0.068732247	0.579879602
공기업16	1	0.079248374	0.123191261	-0.043942886	0.580986163
공기업17	0	0.120547566	0.088918451	0.031629116	0.133581907
공기업18	0	0.126709755	0.100633199	0.026076556	0.163994599
공기업19	0	-0.081205443	0.021354315	-0.102559758	0.301459596
공기업20	0	0.127398270	0.138267635	-0.010869364	0.136179769
공기업21	0	0.146907111	0.341154261	-0.194247150	0.175713418
공기업22	0	0.050126642	0.163478935	-0.113352293	0.542405340
공기업23	1	0.045910015	0.136336224	-0.090426209	0.432730280
공기업24	1	0.085610118	0.140835325	-0.055225207	0.434816043
공기업25	1	0.059213148	0.143019285	-0.083806137	0.447519098
공기업26	1	0.090489510	0.158242894	-0.067753384	0.519685157

□ 공기업 재무건전성 예측모델 설계에 사용한 공기업 재무자료 세부항목 (10/10)

명칭	전기자산회전율	자산회전율의변화	이자보상배율	자산규모의변화	
공기업1	0.512162956	0.062387215	1.246390292	-0.075249516	
공기업2	0.340930837	-0.008280424	2.767816743	0.021978568	
공기업3	0.219852175	-0.005418213	54.880865425	0.120681718	
공기업4	0.189363158	0.006964233	0.0001	0.029332564	
공기업5	0.062757036	-0.002147538	3.081637052	0.004486285	
공기업6	0.048328517	0.003029257	1.706648641	0.020496366	
공기업7	1.226115748	0.019069311	10.259389407	0.009070328	
공기업8	0.570954038	-0.045262561	38.376703262	0.011175514	
공기업9	3.093925133	-0.036194063	0.0001	0.028552072	
공기업10	0.073008205	0.096727072	-69.868134218	-0.050245405	
공기업11	0.211626945	-0.031465348	-1.681982503	0.084983016	
공기업12	0.107367831	0.003918532	0.438561002	-0.124087938	
공기업13	0.344744413	-0.006770119	2.666571288	0.108399826	
공기업14	0.146789122	-0.021442559	137.882240075	0.052347927	
공기업15	0.592228975	-0.012349374	15.476612888	0.163943131	
공기업16	0.594858185	-0.013872022	0.0001	0.021789061	
공기업17	0.139246636	-0.005664729	1.216972270	0.024981110	
공기업18	0.181584861	-0.017590262	1.629111189	0.027590322	
공기업19	0.310181935	-0.008722339	-1.197774470	0.071161473	
공기업20	0.134231416	0.001948353	21.587509080	0.007867765	
공기업21	0.216295520	-0.040582103	2.815448522	0.040839063	
공기업22	0.529647889	0.012757451	2.948343116	0.015463553	
공기업23	0.434850620	-0.002120340	2.569588809	0.082010498	
공기업24	0.439954457	-0.005138415	2.954295572	-0.016933154	
공기업25	0.436589017	0.010930081	2.373343108	-0.003921650	
공기업26	0.473885320	0.045799838	4.262564774	-0.013199999	

ABSTRACT

A Empirical Study on the Financial Stability Prediction Model of South Korea's Public Enterprises with Machine Learning Techniques

Yoon, Hye-Ran
Major in Smart Convergence Consulting
Dept. of Smart Convergence Consulting
The Graduate School
Hansung University

The purpose of this study is to find out how to apply the machine learning technique in the consulting field through the empirical study on the financial stability prediction model of South Korea's Public Enterprises with four machine learning techniques, Random Forest, XGBoost, LightGBM, DNN.

In just two to three years, the rapid growth of the big data market through the digitization of data and the development of computer performance has led to the rapid growth of public interest in the use of machine learning along with the advances in machine learning Techniques(Korea IR Council, 2019).

Traditional statistical analytics techniques focus on mathematical models that explore the meaning of the entire population through sample data, but machine learning focuses on building models that find important patterns and

rules based on data and then perform decision-making support and prediction. Now decision-making systems based on data rather than human intuition are becoming more important, and management prediction based on data is considered most important for the survival and development of companies(Yang, 2017).

Global leading companies, Google, Amazon, and Microsoft are transforming all business capabilities into data and artificial intelligence. Also in South Korea, the government plans to build big data platforms and centers for collecting and providing data by public and private sectors for the use of artificial intelligence and is working various support projects for helping SMEs and venture companies develop new services using data.

As interest and demand for machine learning grow in these diverse fields, consultants also need to understand and find applicability about new methodology named machine learning. the studies with machine learning techniques have been mainly conducted in the fields of statistics, medicine, engineering, and natural sciences, but recently have also started to be conducted in the fields of the social sciences(Choi, Min, 2018).

From last year to this year, there have been continuous reports that the financial stability of South Korea's major public institutions was deteriorating. On the other hand, the studies on improvement of the financial stability of South Korea's public institutions has been continued since 2012, but no study on the prediction of public institution's management performance or financial stability with machine learning techniques has been found. Thus in this study, I designed the financial stability prediction model of public enterprises.

To design the financial stability prediction model of public enterprises, this study used financial data of 26 public enterprises, which were selected for seven consecutive years from 2011 to 2017 when K-IFRS was introduced. In this study, the data were divided into five-year units and designed the 2016 and 2017 financial stability prediction model of public enterprises with four

machine learning techniques.

The analysis results of this study are as follows. First, as a result of the prediction model design, it was possible to make the prediction model of the public enterprises' financial stability with Random Forest and XGboost among four machine learning techniques. In model-specific prediction, both models showed over about 85% prediction accuracy, and the prediction model with XGBoost was higher than the Random Forest. However, with LightGBM and DNN, could not make the prediction model. It is confirmed that the number of data is so small that the over-fitting caused by excessive learning, and so the prediction accuracy is lower due to the large error when predicting the actual data.

In year prediction, the prediction accuracy of the 2017' model was lower than that of the 2016' model in both the prediction model of the public enterprises' financial stability with XGboost and Random Forest. This is due to the change of regime in 2017. With the beginning of the Moon Jae-in government in May 2017, there have been many changes in society, including policies. These changes may be the factors that degrade the prediction accuracy through the existing data.

Machine learning techniques provide variables importance. It is evaluated that the importance score of the independent variable is higher, it is an important variable predicting the target variable. In this study, in each prediction model using Random Forest and XGboost, the main variables were identified as 'Ordinary Income - Underlying Asset', 'Operating Income/Revenue', 'Operating Income - Revenue', 'Operating Income' and Operating Cash Flow'. These variables are mainly related to profitability, proved to be major factors in predicting public enterprises' financial stability.

These results show that public enterprises should pursue public interests, but profitability cannot be overlooked in order to improve the financial stability of public enterprises. Currently, South Korea's public

enterprises' debt continues to increase, but the score of financial management indicators in public enterprises' management performance indicators has continued to decrease, and the proportion of qualitative evaluations has continued to increase. So it is time to reconsider the evaluation indicators of the public enterprises' management performance.

Based on the results of this study, the following are some of the considerations for applying machine learning techniques in consulting fields. First, the amount of data required for machine learning techniques depends on the complexity of the problem and the machine learning technique. The data most enterprises have are unclean and nonlinear, and so much more data is needed to find patterns and rules with these data.

The vast amount of data acquired through the Internet has made it possible to use big data-based machine learning. However, there is no standardized method on the amount of data required for the desired level of a model in machine learning. In addition, as the consulting methodology varies depending on the size of the company or the type of problem the company demand, it is necessary to use appropriate machine learning techniques according to the size and type of data that a company has. Currently Machine learning techniques continue to evolve, it is difficult to find standardized methodologies like traditional statistical analysis. Therefore, it is necessary to accumulate know-how through various trial and error when applying the actual machine learning technique.

In this study, external factors such as the change of regime were not considered in the design of the financial stability prediction model of public enterprises, which can be a major factor affecting actual data prediction. Therefore, when designing the prediction model with the machine learning method, it is necessary to monitor the change in the characteristics of the predicted variable according to the external predicted time and reflect the change or periodically relearn(Žliobaitė, 2010).

On the other hand, through this study, I identified the usefulness of the machine learning technique. First, the prediction model designed with machine learning can easily check the prediction result by using the already designed code when new data of the same type is given. In other words, it was possible to check the result of the 2017' prediction model easily by inputting only code name fitting the year in the code that designed the financial stability prediction model of the 2016' public enterprises. In addition, the variable importance provided by machine learning techniques makes it easy to identify the major variables that contribute to the prediction of target variables. The advantages of this machine learning technique can be a useful tool in consulting.

AI is now a key driver of the Fourth Industrial Revolution, which will bring about changes in the industrial structure as well as social systems, and it is predicted that AI will be the main factor for the company growth(Korea IR Council, 2019). Consultants are experts who provide solutions for enterprises' problem-solving. They should respond sensitively and agilely to changes in the industrial and business environment. Consultants should understand quickly about artificial intelligence which is developing at a rapid pace and prepare how to apply it in consulting.

This study is meaningful in that it tried to find a way to apply the machine learning method in the consulting field through the empirical study on the prediction model with the machine learning method, not the traditional statistical analysis method. It is hoped that this study will be a stepping stone for consultants to apply machine learning techniques in the consulting field with the recognition of the need and importance of machine learning.

【Key Words】 Consulting, public enterprises, financial stability, financial stability score, KJY score, K-IFRS, machine learning, prediction analysis, prediction model, RandomForest, XGBoost, LightGBM, DNN