공학석사 학위논문

조선 생산 리드타임 기준정보 관리를 위한 앙상블 학습 기법 적용 연구

A study on the application of ensemble learning for the management of the shipbuilding production lead time master data

지도교수 남종호

2020년 2월

한국해양대학교 대학원

조선해양시스템공학과

정 주 현

본 논문을 정주현의 공학석사 학위논문으로 인준함.

위원장 박 주 용 (인)

위 원 남 종 호 (인)

위 원 우 종 훈 (인)

2019년 12월

한국해양대학교 대학원

목 차

List of Tables

List of Figures

Study on the application of ensemble learning for the management of the shipbuilding production lead time master data

Jeong, Ju Hyeon

Department of Naval Architecture and Ocean System Engineering
Graduate School of Korea Maritime and Ocean University

Abstract

Building large structures such as ships requires efficient management of production information. In shipyards, enterprise management information that includes production information is called master data, and which includes BOM (bill of material), WBS (work breakdown structure), basic unit and lead time. Master data related to production is closely related to time information. In the shipbuilding industry, however, the high variability of the shipbuilding process has made it difficult to gain the reliability of master data management. Low accuracy of master data can lead to financial losses as well as confusion of work.

To solve these problems, shipyards and related academia have been making various efforts to improve the master data system. However, most of the existing research is focused on traditional engineering perspectives and does not reflect the rapidly changing shipbuilding production environment.

Recently, machine learning methodologies have been widely applied to the manufacturing industry, along with the rapid development of big data related technologies. Machine learning is a correlation analysis technique of vast amounts of data, which is known to be able to overcome the limitations of causation analysis with traditional engineering methodologies. Research is also being conducted in the shipbuilding industry by applying various machine learning algorithms to systematically manage the production master data.

In this paper, I would like to introduce the study of applying ensemble learning, which is known to maximize the performance of machine learning, to the analysis of shipbuilding production master data. In addition, we would like to examine the applicability of production management tasks in actual shipyards by comparing the prediction results of the ensemble learning with the results of applying various learning algorithms and attaching a better performance learning model to the simulation.

KEY WORDS : Shipbuilding, Master data, Machine learning, Ensemble learning, DES simulation

제 1 장 서 론

1.1 연구배경

조선업의 제품은 거대한 3차원 구조물로서 생산 리드타임이 매우 길고 제한된 기간 내에 신규 설계 및 생산이 이루어져야하기 때문에 많은 양의 생산 정보의 생성, 전달, 조정, 변경이 반복된다(Yang, et al., 1992). 그렇기 때문에 선박 또는 해양플랜트와 같은 거대 구조물을 건조하기 위해서는 생산 정보에 대한 효율적인 관리가 필요하다.

조선소에서 납기 준수는 매우 중요하기 때문에, 생산 계획 단계에서 실제 리드타임과 최대한 유사하게 계획 리드타임을 정해야 한다. 하지만 조선업에서의 생산은 선박 건조 공정의 높은 불확실성과 변동성으로 인해 계획과 실제 리드타임에 매우 큰 차이가 존재하고 있어 리드타임 기준 정보 관리의 신뢰성 확보에 어려움을 겪고 있다. 현재는 분기별 피드백이나 연간 사업계획 수정으로 이를 극복하고 있지만, 시간 측면에서 한계가 있는 실정이다.

이러한 문제점을 해결하기 위한 다양한 방법 중 하나로 기계학습 방법론이 주목받고 있다. 기계 학습은 방대한 데이터의 상관관계 분석 기법으로써 기존의 엔지니어링 방법론이 지니고 있는 인과 관계 분석의 한계를 극복할 수 있다고 알려져 있다. 조선소 및 유관 학계에서는 생산 기준 정보에 대한 체계적인 관리를 위해 다양한 기계 학습 알고리즘을 적용한 연구를 진행하고 있으며, 본 연구에서도 기계학습 연구의 일환으로 기계학습의 성능을 극대화 할 수 있다고 알려진 앙상블 학습 기법을 생산 기준정보 분석에 활용하고자 한다. 또한, 선행적으로 진행되었던 기계학습 및 심층학습과 앙상블학습의 예측 결과를 비교하여 가장 좋은 성능을 보이는 학습 모델의 예측 리드타임을 반영한 시뮬레이션을 수행함으로써 실제 조선소에서의 활용성을 검토해 보고자 한다. 이를 통해 조선소에서의 생산 기준 정보를 관리할 수 있는 방안이 마련될 수 있을 것이라 기대하는 바이다.

1.2 관련 연구 동향

1.2.1 조선해양산업 빅데이터 분석 연구 사례

4차 산업 혁명 시대를 맞아 빅데이터 분석을 어떻게 제조 현장에서 활용할 수 있으며 제조 현장에 빅데이터 분석을 적용하기 위해 무엇이 필요한지 명확히 아는 것은 매우 중요하다. 그렇기 때문에 다양한 제조 산업에서는 빅데이터를 도입하기 위한 연구가 활발히 진행 중이다. 이와 더불어 전통적인 제조 산업인 조선업에서도 빅데이터를 도입하기 위한 연구가 활발히 진행되고 있다.

본 논문의 선행연구인 김지혜(2018)에서는 제품의 다양한 속성을 고려한 변동 리드타임을 예측하기 위해 조선소의 다양한 공정 데이터를 수집하였고 공정에 따른 생산 리드타임을 예측하기 위해 다양한 기계학습 및 딥러닝 알고리즘을 적용하였다. 데이터를 분석하기 위해서 R과 Python 언어 등의 오픈소스를 활용하였으며 알고리즘에 따른 리드타임 예측모델을 생성하였다. 또한 분석 알고리즘에 따라 생성된 예측모델의 평가를 위해 여러 가지 평가지표를 활용하였다.

함동균(2016)에서는 의장품 중 후행작업에서 많은 지연이 야기되는 배관재의 제작 공정부터 설치 공정까지의 리드타임을 예측하여 조달관리의 수준을 높이기 위한 연구를 수행하였다. 해당 연구에서는 배관공정의 공급망을 6개의 공정으로 나누어 리드타임을 정의하였으며 이를 예측하기 위하여 SPSS를 활용한 다중선형회귀분석과 PLS 회귀분석을 수행하였다.

이동하 등(2013)에서는 조선 산업에서 블록 조립 작업에 대한 계획 프로세스와 실적 프로세스를 비교하는 방법을 제안하는 연구를 수행하였다. 해당 연구에서 제안한 방법은 계획과 실적 데이터를 기반으로 프로세스 마이닝 기법을 이용하여 프로세스 모델을 도출하고 비교 분석하는 것이다. 이를 통해 형상과 구조가 모두 다른 다양한 조립 블록에 대해서 쉽고 빠르게 프로세스 모델을 정의하고 계획과 실적 관점에서 작업의 특성들을 비교할 수 있다.

김영주 등(2013)에서는 조선 산업의 선박설계 생산성 향상을 위한 선박설계 자동화에

관한 연구를 수행하였다. 해당 연구에서는 선박설계 자동화를 지원하기 위해 선박 설계의 주요 기술과 요구조건을 분석하고 선박설계 자료를 효과적으로 관리하기 위한 빅데이터 기술 및 분석기법을 연구하였으며 분석된 결과를 반영한 선박 설계 자동화 시스템 설계를 제안하였다.

김성훈 등(2016)에서는 대표적인 빅데이터 프레임워크인 하둡(Hadoop)을 활용한 빅데이터랫폼을 제시하고, 이를 FPSO 상부의 중량을 추정하는데 활용함으로써 그 적용 가능성을 확인하였다. 해당 연구에서 제시한 빅데이터 플랫폼은 기존의 대용량 데이터를 하둡 기반의 분산 저장 및 처리구조(HDFS)로 개선한다. 이를 구현하기 위해 R 프로그램과 Rhadoop 패키지를 이용하였고, 그 결과 하둡의 HDFS와 MapReduce 기능을 효과적으로 활용 가능함을 확인하였다.

1.2.2 앙상블학습 적용 연구 사례

김민석 등(2012)에서는 반도체 산업의 품질보증검사 불량 예측에 데이터마이닝 학습방법을 적용하여 정상로트와 불량로트를 예측하는 방법을 제안한다. 수집된 로트 데이터를 사용하여 분류 성능을 최대화 할 수 있는 앙상블 학습을 통해 모형을 구축하고 동일한 현장의 데이터로 성능을 비교했다. 성능평가 결과 배깅과 의사결정나무를 조합한 모형의 우수성을 확인했고 현업 적용가능성도 확인했다.

민성환(2014)에서는 기업의 부도 예측 모형의 성과를 개선하기 위해 사례 선택과 배깅을 연결하는 새로운 모형을 제안하였다. 최적의 사례 선택을 위해 유전자 알고리즘이 사용되었으며, 이를 통해 최적의 사례 선택 조합을 찾고 이 결과를 배깅 앙상블 모형에 전달하여 새로운 형태의 배깅 앙상블 모형을 구성하게 된다. 실제 기업데이터를 사용해 실험한 결과 해당 연구에서 제안한 새로운 형태의 모형이 가장 좋은 성과를 보임을 알 수 있었다.

이상헌 등(2014)에서는 운전자의 차선 변경 의도를 예측하기 위해 앙상블 기법을 활용하였다. 해당 연구에서는 adaboost 앙상블 기법을 도입했으며 실험데이터를 기존의 단일 분류기와 앙상블 기법 각각에 대해 적용하여 성능 향상 효과를 관찰하였다. 그 결과 앙상블 기법의 성능이 기존의 단일 분류기에 비해 월등히 우수함을 확인할 수

있었다.

박선 등(2012) 에서는 의사결정트리를 기본 분류기로 이용한 bagging 앙상블 학습과 adaboost 앙상블 학습을 이용하여 미래에 발생할 적조의 예측 정확도를 향상시키는 방법을 제안하였다. 통영지역의 2002년부터 2007년 동안 발생한 적조 정보와 같은 지역의 해양 환경정보를 학습하였으며, 이후 4년간의 정보를 이용하여서 제안 방법을 평가하였다. 평가결과, 의사결정트리를 기본 분류기로 이용한 bagging 앙상블 학습 방법이 적조 발생의 예측 성능을 향상시켰다.

현재까지 조선업에서 빅데이터가 실질적으로 적용된 사례는 드물지만, 이를 대상으로 한 연구 사례는 계속해서 늘어나고 있다. 또한 빅데이터 분석의 한 방법론인 앙상블 학습이 다양한 분야에 연구되고 있으며, 연구 결과를 통해 기존의 단일 학습기에 비해 우수한 성능을 보이는 것을 확인할 수 있었다. 하지만 앙상블 학습이 조선업에 적용된 사례는 찾아볼 수 없었으며 다양한 연구를 통해 빅데이터 기술은 확보되었지만 적용대상 즉, 실제 조선소의 데이터가 없는 경우가 존재하는 실정이다.

1.3 연구 목적

본 연구에서는 기존의 인과관계 분석의 한계에서 벗어나 현장 상황이나 제품, 공정 정보 등 다양한 속성을 반영한 예측 리드타임을 통해 생산 리드타임 기준정보의 체계적인 관리를 가능하게 하고자 한다. 본 연구에서는 Python 언어를 활용하여 데이터 분석을 수행하였으며 선행적으로 진행되었던 기계학습, 심층학습에 이어 앙상블 학습으로 기계학습 방법론을 확장하였고, 실제 조선소의 데이터에 앙상블학습을 적용하여 생산 리드타임을 예측하고자 한다.

이를 정리하자면 기존의 조선소에서는 선박 건조공정의 높은 불확실성과 변동성으로 인해 계획과 실제 리드타임에 큰 차이가 발생하게 되고 이로 인한 비용적 손실이 발생할 수 있다. 따라서 Fig. 1과 같이 빅데이터 분석을 통해 예측된 생산 리드타임을 반영하여 생산 계획을 수립함으로써 기존의 문제점을 해결해 보고자 한다.

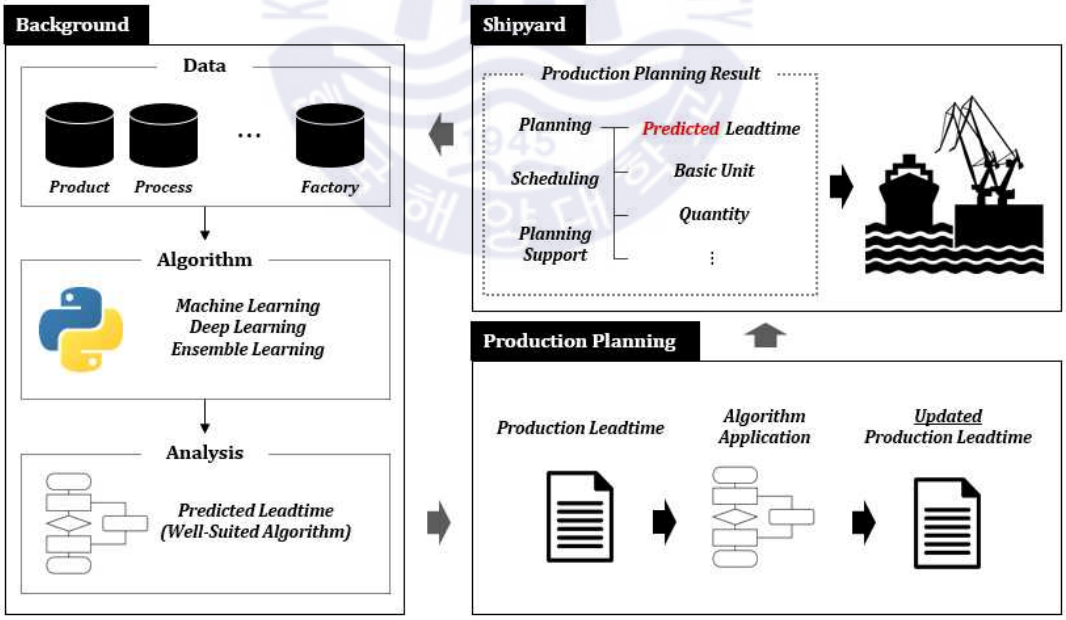


Fig. 1 Application of production lead time prediction model

제 2 장 적용개념

2.1 분석 알고리즘 소개

본 연구에서 분석에 활용되는 알고리즘은 크게 기계학습, 심층학습, 앙상블학습이다. 심층학습과 앙상블학습은 기계학습의 한 방법론에 해당하지만 알고리즘을 확장시켜 나가는 과정을 보이기 위해서 편의상 구분해서 사용하도록 한다.

2.1.1 기계학습 알고리즘

기계학습은 인공지능의 한 분야로 컴퓨터가 스스로 학습할 수 있도록 하는 알고리즘과 기술을 개발하는 분야이다. 기계학습은 데이터를 이용하여 모델을 만드는 것이 목적이기 때문에 적절한 입력 데이터 선정뿐만 아니라 문제에 적합한 알고리즘을 선택하는 것 또한 중요하다. 기계학습 알고리즘은 크게 지도 학습, 비지도 학습, 강화 학습으로 분류할 수 있다.

지도 학습이란 훈련 데이터에 레이블(label)이 있는 경우, 훈련 데이터로부터 하나의 함수를 유추해 내기 위한 학습을 의미한다. 따라서 지도 학습은 명확한 입력 값과 출력 값이 존재하며, 주어진 데이터를 정해진 범주에 따라 나누는 분류와 연속된 값을 예측하는 회귀로 구분할 수 있다. 비지도 학습이란 레이블이 없는 훈련 데이터로부터 어떠한 지식을 추출하는 것을 의미한다. 즉 사람이 개입하지 않고 컴퓨터 스스로 데이터를 훈련하는 것이다. 강화 학습이란 기계학습이 다루는 문제들 중 하나로 어떤 환경 안에서 정의된 에이전트가 현재의 상태를 인식하여, 선택 가능한 행동들 중 보상을 최대화하는 행동 혹은 행동순서를 선택하는 방법이다.

본 연구의 목적은 리드타임이라는 기준 정보를 개선하기 위한 예측모델을 개발하는 것이다. 따라서 조선소의 공정정보가 입력 값, 예측 대상인 리드타임이 출력 값으로 정의되기 때문에 본 연구에서의 기계학습 알고리즘은 지도학습 알고리즘으로 한정된다. 그리고 이러한 지도학습 알고리즘 중에서도 수치 예측에 적합하다고 알려진 회귀분석(regression analysis)과 의사결정나무(decision tree)를 활용하고자 한다.

2.1.1.1 회귀분석(regression analysis)

회귀분석은 하나 혹은 그 이상의 독립변수들이 종속변수에 미치는 영향을 추정하는 통계기법으로 변수들 사이의 인과관계를 규명하고자 하는 분석 방법이기 때문에 변수의 역할 설정이 중요하다. 회귀분석에서 다른 변수에 영향을 주는 원인에 해당하는 변수를 독립변수(independent variable) 또는 설명변수(explanatory variable)라고 하며, 영향을 받는 결과에 해당하는 변수를 종속변수(dependent variable) 또는 반응변수(response variable)라고 한다. 회귀분석은 독립변수와 종속변수 사이의 구체적인 함수식을 찾아내고, 독립변수로부터 종속변수를 예측하는데 그 목적이 있다. 독립변수와 종속변수가 각각 1개일 때의 분석을 단순회귀분석(simple regression analysis), 종속변수가 1개이면서 독립변수가 2개 이상일 때의 분석을 다중회귀분석(multi regression analysis)이라고 한다. 본 연구에서는 리드타임이라는 하나의 종속변수와 여러 개의 독립변수 사이의 관계를 분석하는 것이기 때문에 다중 회귀분석을 활용하고자 한다.

2.1.1.2 의사결정나무(decision tree)

의사결정 규칙을 나무구조로 도표화하여 분류와 예측을 수행하는 분석방법으로 분류 또는 예측이 나무구조에 의한 추론규칙에 의해서 표현되기 때문에 다른 방법들에 비해 그 과정을 쉽게 이해하고 설명할 수 있다. 의사결정나무분석은 일반적으로 분석의 목적과 자료구조에 따라서 적절한 분리기준과 정지규칙을 지정하여 의사결정나무를 얻는 의사결정나무의 형성, 분류오류를 크게 할 위험이 높거나 부적절한 추론규칙을 가지고 있는 가지를 제거하는 가지치기, 이익도표나 위험도표 또는 검증용 자료에 의한 교차 타당성 등을 이용하여 의사결정나무를 평가하는 타당성 평가, 의사결정나무를 해석하고 예측모형을 설정하는 해석 및 예측 순서로 진행된다. 의사결정나무는 대표적인 분류 모델이지만 대상 변수가 범주형이면 분류 트리, 연속형이면 회귀 트리로 분류된다. 본 연구에서는 대상 변수가 연속형인 리드타임이기 때문에 회귀 트리로 분석을 수행한다.

2.1.2 심층학습 알고리즘

심층학습이란 다수의 신경층을 가진 인공신경망을 사용하여 기계학습을 수행하는 것으로 딥러닝 이라고도 부른다. 따라서 심층학습은 기계학습과 전혀 다른 개념이 아니라 기계학습의 한 종류라고 할 수 있다.

기존의 기계학습과의 차이점이 있다면, 기계학습은 기계가 학습하기 위해 주어진 데이터로부터 특징을 추출하는 과정에서 사람이 개입하지만 심층학습은 주어진 데이터를 그대로 활용하여 컴퓨터가 스스로 학습한다는 점이다. 사람이 생각한 특징을 훈련하는 것이 아니라 데이터 자체에서 중요한 특징을 기계 스스로 학습하여 사람이 개입함으로써 생길 수 있는 오류를 줄일 수 있다. 또한 다수의 신경층을 이용하는 접근은 비선형 문제, 계층의 수에 따른 가중치 수의 한계, 과적합 등의 문제점으로 인해 활용되지 않았다. 그러나 이러한 문제점은 컴퓨터의 계산 성능과 알고리즘의 발달로 인해 다층 신경망의 효용성이 밝혀지게 되면서 심층학습 기술을 현재 인공지능 분야에서 널리 활용되고 있다(Kim, 2016). 본 연구에서는 수치예측에 많이 활용되는 심층학습 알고리즘 중에서도 가장 기본적인 다층 퍼셉트론(multi-layer perceptron)을 활용하고자 한다.

2.1.2.1 다층 퍼셉트론(multi-layer perceptron)

다층 퍼셉트론은 퍼셉트론이 여러 층으로 이루어져있는 형태로써, 이를 이해하기 위해서는 먼저 퍼셉트론을 이해해야 한다. 퍼셉트론은 다수의 입력 신호를 받아서 하나의 신호를 출력하는데, 이는 인간의 신경세포인 뉴런을 매우 단순하게 모사하여 계산 가능한 형태로 만든 알고리즘이라고 할 수 있다. 다수의 입력 신호를 받았을 때, 퍼셉트론은 각 입력신호의 세기에 따라 다른 가중치를 부여한다. 이러한 가중치와 입력값을 곱하여 모두 합한 값이 임계치보다 크면 활성화되어 1을 출력하고 활성화되지 않으면 결과 값으로 0을 출력한다. 이를 그림으로 나타내면 Fig. 2와 같다.

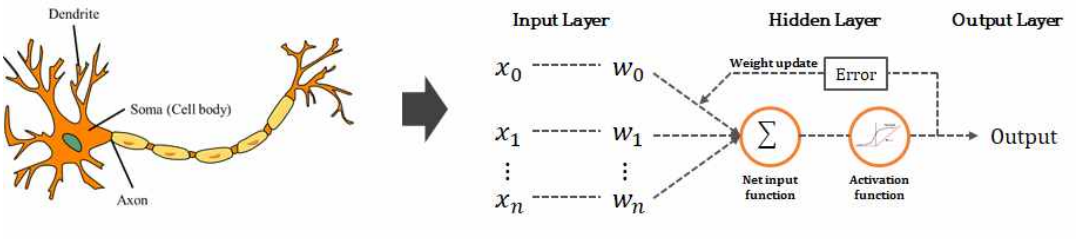


Fig. 2 Conceptual diagram of perceptron

기본적인 퍼셉트론은 데이터의 입력층과 출력층만 있는 구조로 단층 퍼셉트론이라고도 하는데 이는 활성 함수가 1개 밖에 없는 구조이기 때문에 비선형적으로 분리되는 데이터에 대해서는 제대로 된 학습이 불가능하다는 문제점이 있다. 따라서 이를 극복하기 위한 방안으로 다층 퍼셉트론이 고안되었다. 다층 퍼셉트론의 구조는 Fig. 3과 같고, 이는 입력층과 출력층 사이에 하나 이상의 중간층이 존재하는 신경망으로, 이 때 입력층과 출력층 사이의 중간층을 은닉층이라 부른다. 네트워크는 입력층, 은닉층, 출력층 방향으로 연결되어 있으며, 각 층 내의 연결과 출력층에서 입력층으로의 직접적인 연결은 존재하지 않는 전방향 네트워크(feedforward network)이다.

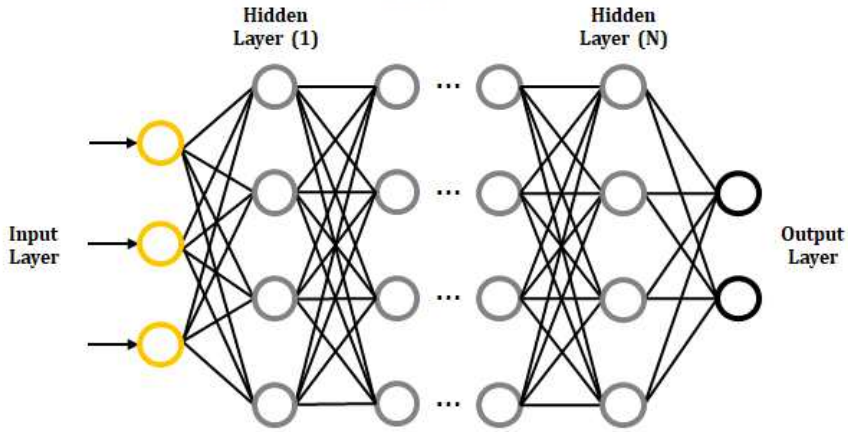


Fig. 3 Structure of multi-layer perceptron

2.1.3 앙상블학습 알고리즘

앙상블학습이란 여러 개의 단일 학습 모델을 학습하고 그것들의 예측을 결합함으로써 새로운 가설을 학습하는 방법으로, 다양한 학습 모델의 예측 결과를 결합함으로써 단일 학습 모델보다 신뢰성이 높은 예측 값을 얻는 것이 앙상블 학습의 목표이다(Lee & Yang, 2013). 앙상블 학습의 구조를 간략한 그림으로 나타내면 Fig. 4와 같으며, 앙상블학습의 사용은 학습 모델의 성능을 분산시키기 때문에 과적합(overfitting)이 감소하여 성능을 향상시킬 수 있다는 장점이 있다. 여기서 과적합이란 학습 데이터를 과하게 학습함으로써 학습데이터 셋 안에서는 일정 수준 이상의 예측 정확도를 보이지만 새로운 데이터에 적용하면 예측 정확도가 감소하는 것을 의미한다. 앙상블 학습은 크게 배깅(bagging), 부스팅(boosting), 스태킹(stacking)으로 분류할 수 있다. 본 연구에서는 이에 해당하는 다양한 알고리즘으로 분석을 수행하였다.

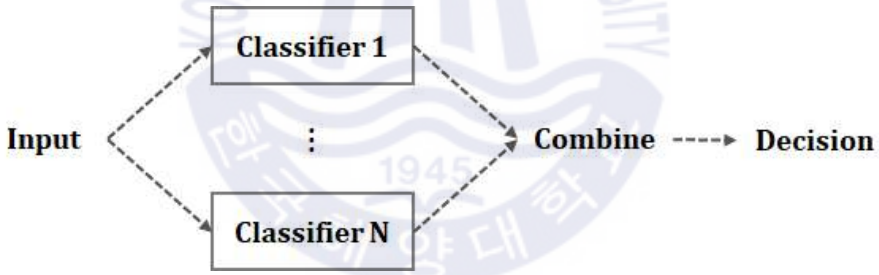


Fig. 4 Structure of ensemble learning

2.1.3.1 배깅(bagging)

bootstrap aggregating의 줄임말로 Fig. 5와 같은 구조를 갖는다. 배깅은 하나의 알고리즘을 사용하지만 학습 데이터 셋을 무작위로 나누어서 학습 모델을 각각 다르게 학습시키고 생성된 학습 모델들의 결과를 종합하여 의사결정을 내리는 방법이다. 이 때, 학습 데이터 셋에서 중복을 허용해서 샘플링하는 복원랜덤샘플링을 수행하게 된다. 배깅은 예측 모형의 변동성이 큰 경우 예측 모형의 변동성을 감소시키기 위해 사용하는 것으로, 여러 번의 복원 샘플링을 통해 예측 모형의 분산을 줄여줌으로써 예측력을 향상시키게 된다.

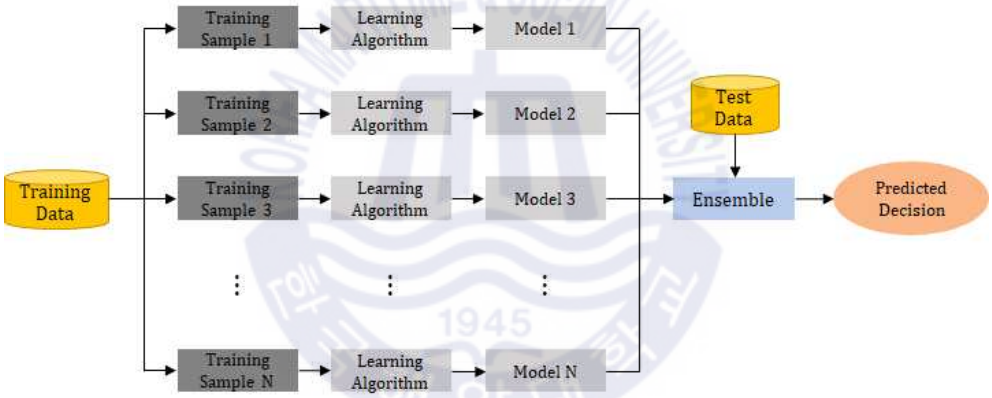


Fig. 5 Flow diagram of bagging

①랜덤 포레스트(random forest)

랜덤 포레스트란 기계학습의 의사결정나무를 개선한 알고리즘이다. 다수의 의사결정나무를 결합하여 하나의 학습 모델을 생성하는 방법으로, 주로 회귀 분석, 분류 등에 사용된다. 랜덤 포레스트는 각각의 의사결정나무를 구성할 때, 훈련 데이터 셋에서 중복을 허용해서 샘플링하는 배깅 방법을 사용해 변수를 선택하게 된다. 매번 다른 독립변수를 갖기 때문에 각각의 단위 모델에 큰 차이가 존재하고, 각 모델들의 예측 결과를 종합함으로써 단일 모델보다 신뢰성이 높은 예측 값을 획득할 수 있다. 기존의

앙상블 모형과 달리 임의성을 관측치 뿐만 아니라 변수에도 적용했기 때문에 앙상블 학습이 갖는 장점을 극대화하여 예측 및 분류 정확도를 기존의 방법보다 개선하며 안정성을 얻게 된다. 랜덤 포레스트는 B개의 의사결정나무 모델을 결합하여 최종 앙상블 학습 모델인 $C^*(x)$를 구축하며, 분석 대상에 따라 최종 학습 모델을 구축하는 방법이 다르다.

회귀 모형인 경우, 식 (1)과 같이 각각의 의사결정나무 예측 값의 평균을 취함으로써 최종 학습 모델을 구축하게 된다.

$$C^*(x) = \sum_{b=1}^{B} C_b(x)/B \tag{1}$$

분류 모형인 경우, 식 (2)와 같이 투표를 통해 가장 많이 선택된 클래스를 반환함으로써 최종 학습 모델을 구축하게 된다.

$$C^*(x) = \mathrm{arg}max_y \sum_{b=1}^{B} I[C_b(x) = y] \tag{2}$$

② 엑스트라 트리(extra-trees)

extremely randomized trees의 줄임말로 극단적으로 무작위한 트리의 랜덤 포레스트를 의미한다. 엑스트라 트리는 트리를 더욱 무작위하게 만들기 위해 최적의 임계값을 찾는 대신 후보 특성을 무작위로 분할한 다음 최적의 분할을 선택한다. 쉽게 말해 랜덤 포레스트와는 다른 방식으로 모델에 무작위성을 주입한다고 볼 수 있다. 모든 노드에서 특성마다 가장 최적의 임계값을 찾는 것이 트리 알고리즘에서 가장 시간이 많이 소요되는 작업 중 하나이므로 일반적인 랜덤 포레스트보다 엑스트라 트리가 훨씬 빠르다. 그러나 엑스트라 트리의 무작위 분할 때문에 일반화 성능을 높이려면 종종 많은 트리를 만들어야 한다는 단점이 있다.

2.1.3.2 부스팅(boosting)

부스팅이란 배깅과 마찬가지로 복원랜덤샘플링을 통해 학습된 여러 개의 약한 학습 모델로 강한 학습 모델을 생성하는 방법으로 Fig. 6과 같은 구조를 갖는다. 일반적인 분류 문제와는 달리 잘 분류되지 못한 개체들에 집중하여 새로운 분류 규칙을 만든다. 부스팅의 가장 큰 특징은 다음 단계의 약한 학습 모델이 이전 단계의 약한 학습 모델의 영향을 받는다는 점으로 이전의 학습 모델보다 더 나은 학습 모델을 만드는 방향으로 각 학습 모델에 가중치를 부여하게 된다. 이 때, 오답에 대해 높은 가중치를 부여하고 정답에 대해 낮은 가중치를 부여하기 때문에 오답에 더욱 집중할 수 있게 된다.

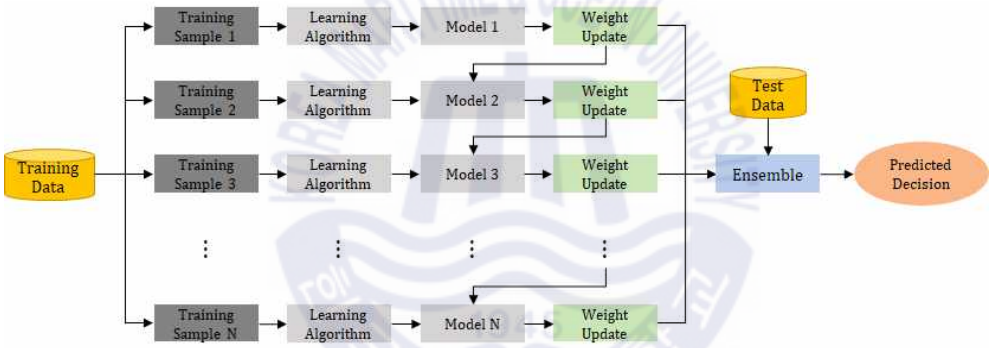


Fig. 6 Flow diagram of boosting

① 아다부스트(adaboost)

adaptive boosting의 줄임말로 이전 모델이 과소적합했던 훈련 샘플의 가중치를 높임으로써 다음 모델을 훈련시키는 알고리즘이다. 이로써 새로운 학습 모델은 학습하기 어려운 샘플에 점점 더 맞춰지게 된다. 예를 들어 먼저 훈련 데이터 셋을 통해 약한 학습 모델을 만든 후, 잘못 분류된 훈련 샘플의 가중치를 상대적으로 높이게 된다. 다음 학습 모델은 업데이트된 가중치를 반영해 새로운 모델을 만들게 되며 이를 반복함으로써 결과적으로 강한 학습 모델이 된다.

② 그래디언트 부스팅(gradient boosting)

그래디언트 부스팅은 아다부스트처럼 학습된 모델의 오차를 보완하는 방향으로 모델을 추가해준다. 하지만, 그래디언트 부스팅은 아다부스트처럼 학습단계마다 데이터 샘플의 가중치를 업데이트 해주는 것이 아니라 전 단계의 학습 모델에서의 잔여 오차에 대해 다음 모델을 학습시키는 알고리즘이다. 그래디언트 부스팅은 주로 의사결정나무 알고리즘과 함께 사용하며 이를 GBRT(gradient boosting regression tree)라고 한다. GBRT는 회귀와 분류 문제에 모두 사용할 수 있으며 보통 하나에서 다섯 개 정도의 깊지 않은 트리를 사용하기 때문에 메모리를 적게 사용하고 예측도 빠르다.

③ xg boost

extreme gradient boost의 줄임말로 불필요한 기능이 거의 없는 오직 속도와 모델 성능에만 초점이 맞추어진 알고리즘이다. xg boost는 트리를 만들 때 CART(classification and regression trees)라 불리는 앙상블 모형을 사용하며 CART는 분류 및 회귀 나무들로 이루어진 트리 기반 앙상블 모형을 의미한다. 이후 트리 부스팅을 사용하여 각 학습 단계마다 가중치를 최적화한다. xg boost는 유연하며 병렬 처리가 가능하기 때문에 학습과 분류가 빠르고, greedy algorithm을 사용한 자동 가지치기가 가능하여 과적합이 잘 일어나지 않는다.

2.1.3.3 스태킹(stacking)

stacked generalization의 줄임말로 Fig. 7과 같은 구조를 갖는다. 스태킹은 다양한 알고리즘을 함께 사용하며, 서로 다른 여러 개의 모델을 조합하여 새로운 메타 모델을 생성하는 형태의 앙상블 모델이다. 즉, 서로 다른 모델들을 조합해서 최고의 성능을 내는 새로운 모델을 생성하는 알고리즘이다. 이러한 조합을 통해 서로의 장점은 취하고 약점을 보완할 수 있다.

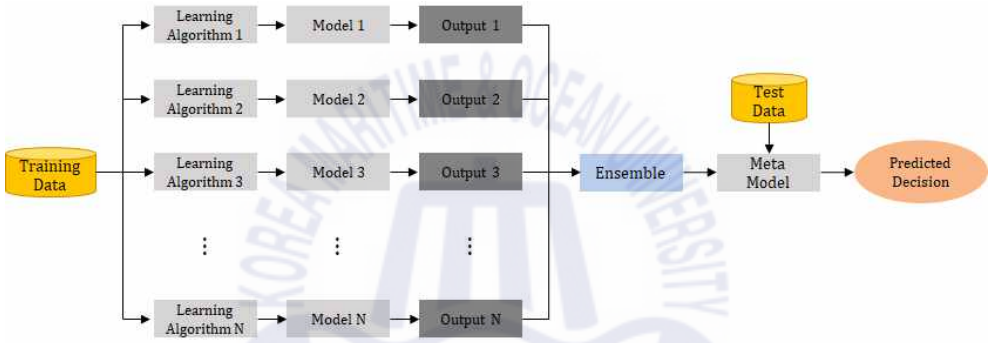


Fig. 7 Flow diagram of stacking

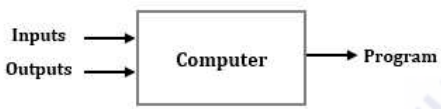
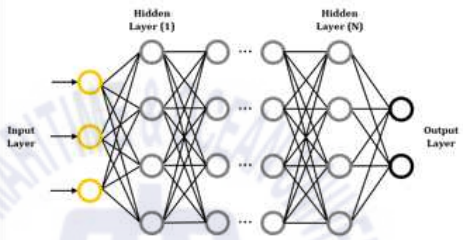
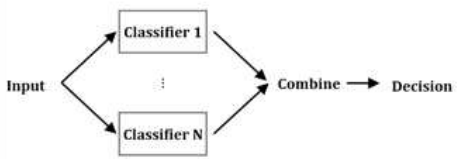
<table>
<tr><th></th><th>Machine Learning</th><th>Deep Learning</th><th>Ensemble Learning</th></tr>
<tr><td>원리</td><td> 
</td><td> 
</td><td> 
</td></tr>
<tr><td>정의</td><td>컴퓨터가 스스로 학습할 수 있도록 하는
알고리즘과 기술을 개발하는 방법</td><td>다수의 신경층을 가진 인공신경망을 사용하여
기계학습을 수행하는 방법</td><td>여러 개의 단일 학습모델을 결합함으로써
더 좋은 성능을 얻고자 하는 방법</td></tr>
<tr><td rowspan="5">적용
알고리즘</td><td rowspan="2">다중 선형 회귀 분석
(multiple linear regression analysis)</td><td rowspan="4">다층 퍼셉트론
(multi-layer perceptron)</td><td rowspan="2">배깅 (bagging)</td><td>random forest</td></tr>
<tr><td>extra trees</td></tr>
<tr><td rowspan="3">의사결정나무 (decision tree)</td><td rowspan="3">부스팅 (boosting)</td><td>ada boost</td></tr>
<tr><td>gradient boosting</td></tr>
<tr><td>xg boost</td></tr>
<tr><td colspan="3">스태킹 (stacking)</td></tr>
</table>

Fig. 8 Summary of analysis algorithm

2.2 데이터 분석 과정

데이터 분석 과정은 Fig. 9와 같으며 크게 데이터 수집, 데이터 전처리, 모델 학습, 학습 모델 성능 확인까지 4단계로 정의하였다. 전반적인 과정을 간단히 설명하자면, 분석 목적에 맞는 데이터를 수집하고 모델 학습에 적합한 형태로 데이터를 전처리 해준다. 다음 단계에서는 주어진 문제와 데이터에 맞는 적절한 알고리즘을 선택하여 모델을 학습한다. 마지막으로 학습 모델의 성능을 평가하고 새로운 데이터에 대한 예측을 하게 된다.

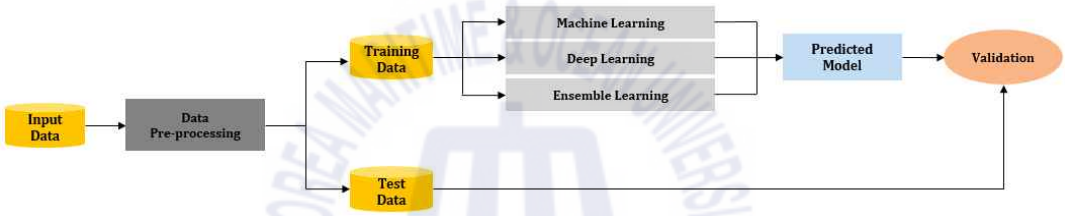


Fig. 9 Data analysis process

데이터 분석 과정의 단계별 상세내용은 다음과 같다.

2.2.1 데이터 수집

데이터 수집 단계에서는 문제 정의 후 이에 맞는 분석 목적에 따라 데이터를 수집한다. 이 과정에서 데이터 분석의 목적을 분명히 하는 것이 매우 중요하다. 그 이유는 데이터 분석의 목적이 분명해야 수집해야 할 데이터가 어떤 정보를 담고 있어야 하는지 구체적으로 검토하고 평가할 수 있기 때문이다.

데이터 수집 단계에서 유의해야 할 점을 크게 일관성과 무작위성으로 정리할 수 있다. 데이터 수집 단계에서 일관성은 매우 중요한데, 이는 데이터를 모으는 과정에서 데이터 수집 방법이 바뀐다면 결과의 품질을 보장할 수 없기 때문이다. 또한 많은 수집이 전체 현상에서 추출한 표본을 대상으로 이루어지는데, 분석 과정에서의 오류를 최소화하기 위해서는 표본 선정의 무작위성(randomness)이 보장되어야 한다. 왜냐하면 데이터를 수집하는 방식에 체계적 편향이 있을 경우 산정된 결과 값은 완전히 다른 수치가 나올 수도 있기 때문이다.

데이터를 수집하는 방법은 여러 가지가 있다. 웹사이트에 있는 자료들을 긁어오기 위해서는 웹 크롤링(web crawling)을 할 수도 있고, 자신이 운영하는 서비스에서 유저들의 행동 데이터를 수집하기 위해 로그를 남길 수도 있다. 또는 이미 데이터베이스에 데이터가 쌓여 있는 경우, 데이터 수집 과정은 간단하게 데이터베이스나 데이터 파일에서 데이터를 불러오는 것으로 충분할 수도 있다.

2.2.2 데이터 전처리

수집된 데이터를 정확하게 분석하기 위해서는 이에 적합한 정확한 데이터가 필요하며, 이를 확보하기 위해서는 충분한 데이터 전처리 과정이 우선적으로 진행되어야 한다. 고도의 분석 기술과 올바를 절차를 따르더라도 정제된 데이터가 확보되지 않으면 왜곡된 분석결과가 나올 수 있으며, 이는 분석에 대한 신뢰도를 떨어뜨릴 뿐만 아니라 동시에 잘못된 의사결정을 유도함으로써 경제적, 사회적 비용을 발생시킬 수 있다.

데이터 전처리를 수행하기에 앞서 분석 목적에 따라 독립변수와 종속변수를 정의하고, 각각의 변수와 데이터 유형을 확인해야 한다. 이는 데이터 유형에 따라 적용되는 분석 알고리즘이 다르며 예측의 결과에도 영향을 미치기 때문에 가장 우선적으로 진행되어야 한다. 그리고 각각의 변수들을 탐색하여 결측값이나 이상치가 있는지 파악한 후 제거함으로써 분석결과의 오류를 사전에 방지하고자 한다. 마지막으로 분석 알고리즘을 적용하기에 적합한 데이터 유형으로 변경함으로써 데이터 전처리를 마무리한다. 본 연구에서는 데이터 셋 확인 단계를 거친 후, 상관분석과 분산분석, 결측값 처리, 이상치 처리, 원-핫 인코딩 순서로 데이터 전처리를 수행한다.

2.2.2.1 상관분석, 분산분석

변수 간 높은 상관 계수가 존재한다는 것은 두 변수가 같이 커지거나 작아지는 경향이 있다는 의미이다. 기계학습 모델에서는 상관 계수가 큰 예측 변수들이 있을 경우 성능이 떨어지거나 모델이 불안정해진다. 또한 기계학습이란 결국 모델의 파라미터를 측정하는 작업인데, 상관 계수가 높은 변수가 여럿 존재하면 파라미터 수가 불필요하게 증가하여 차원의 저주에 빠질 우려가 있다. 그렇기 때문에 상관관계가 높은 변수들이 있다면 이들을 찾아내어 처리해 주어야 한다. 따라서 본 연구에서는 데이터 분석에 활용할 변수를 선별하기 위해 연속형 변수에는 상관분석, 범주형 변수에는 분산분석을 활용하고자 한다.

상관분석(correlation analysis)은 연속형 변수로 측정된 두 변수 간에 어떤 선형적 관계를 갖고 있는지를 분석하는 방법으로 두 변수의 관계가 갖는 강도를 상관관계라고 한다. 상관관계의 정도를 파악하는 상관계수는 -1에서 1사이의 값을 가지며, 변수와의 방향은 Fig. 10과 같이 음의 상관관계일 경우 (-), 양의 상관관계일 경우 (+)로 표현한다. 이 때, 상관관계는 두 변수간의 연관된 정도를 나타낼 뿐 인과관계를 설명하지 않는다는 점을 유의해야 한다. 본 연구에서는 두 연속형 변수간의 관련성을 구하기 위해 보편적으로 사용되는 pearson 상관계수를 활용하고자 하며, 상관계수 절댓값 0.65를 기준으로 보다 큰 값을 갖는 경우 변수에서 제외한다.

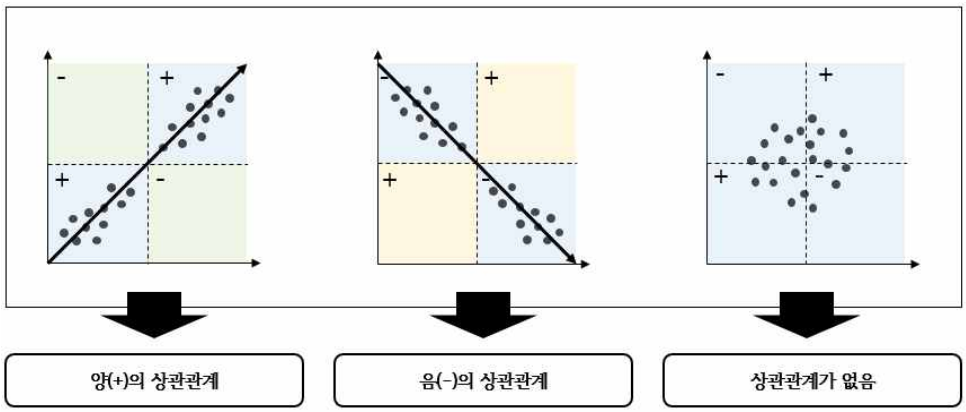


Fig. 10 Correlation analysis

분산분석(ANOVA, analysis of variable)은 관측 자료가 3개 이상의 집단으로 구성된 경우, 집단 간 평균 즉 퍼진 정도를 비교하여 통계적 유의성을 검증하는 방법으로 독립변수는 범주형 변수, 종속변수는 연속형 변수여야 한다. 보통 집단이 2개인 경우에는 t검정을 사용하지만 집단이 3개 이상인 경우라면 t검정을 사용하기에 한계가 존재하기 때문에 분산분석을 활용한다. 분산분석은 종속변수가 1개인 경우 독립변수의 수에 따라 구분할 수 있는데, 독립변수가 1개인 일원분산분석, 독립변수가 2개인 경우를 이원분산분석, 독립변수가 3개 이상인 경우를 다원분산분석이라 한다. 일반적으로 0.01, 0.05, 0.1을 유의수준으로 설정하며, 본 연구에서는 이 중에서 평균에 해당하는 0.05를 기준으로 보다 큰 값을 갖는 경우 변수에서 제외한다.

2.2.2.2 결측값 처리

결측값(missing value)는 말 그대로 데이터에 값이 없는 것을 뜻하며, NA 또는 Null이라고 표현되기도 한다. 결측값이 있는 상태로 모델을 만들게 될 경우 변수간의 관계가 왜곡될 수 있기 때문에 모델의 정확성이 떨어진다. 그렇기 때문에 결측값을 처리하는 과정은 필수적으로 진행되어야 하며, 결측값을 처리하는 방법은 결측값이 발생하는 유형에 따라 달라진다.

① 삭제

결측값이 발생한 모든 관측치를 삭제하거나(전체 삭제) 학습 모델에 포함시킬 변수들 중 결측값이 발생한 모든 관측치를 삭제하는 방법(부분 삭제)이 있다. 전체 삭제는 간편한 반면 관측치가 줄어들어 학습모델의 유효성이 낮아질 수 있고, 부분 삭제는 모델에 따라 변수가 제각각 다르기 때문에 관리 cost가 늘어난다는 단점이 있다. 삭제는 결측값이 무작위로 발생한 경우에 사용한다. 결측값이 무작위로 발생한 것이 아닌데 관측치를 삭제한 데이터를 사용할 경우 오히려 왜곡된 모델이 생성된 수 있다.

② 대체

결측값이 발생한 경우 다른 관측치의 평균, 최빈값, 중간값 등으로 대체할 수 있다. 연속형 변수의 경우 모든 관측치의 평균값 등으로 대체하는 일괄 대체 방법, 범주형 변수의 경우 유사한 유형의 평균값 등으로 대체하는 유사 대체 방법을 적용할 수 있다. 결측값의 발생이 다른 변수와 관계가 있는 경우에는 대체 방법이 유용할 수 있지만, 유사 대체 방법의 경우 유사한 유형을 선택하는 과정에서 개인의 의견이 반영되기 때문에 학습 모델이 왜곡될 가능성이 존재한다.

③ 예측값 삽입

결측값이 없는 관측치를 학습 데이터로 사용해서 결측값을 예측하는 모델을 만들고 이 모델을 통해 결측값이 있는 관측 데이터의 결측값을 예측하는 방법이다. 이를 위해서 주로 regression이나 logistic regression을 사용한다. 대체하는 방법보다는 덜 자의적이지만 결측값이 다양한 변수에서 발생하는 경우 적합한 모델을 만들기 어렵고, 이렇게 만들어진 모델의 예측력이 낮은 경우 사용하기 어렵다는 단점이 있다.

2.2.2.3 이상치 처리

이상치(outlier)란 데이터의 전체적인 패턴에서 벗어난 관측 값으로 잘못된 분석결과를 초래할 수 있는 값을 의미한다. 이상치는 그림을 이용한 탐색을 통해서 발견할 수 있다. 주로 상자그림(box-plot), 히스토그램, 산점도(scatter plot)를 사용하며, 수치적으로는 다음의 기준에 의해 이상치를 찾을 수 있다.

① iqr(interquartile range) rule

iqr rule은 Fig. 11과 같이 box plot으로 데이터의 분포를 시각화함으로써 설명할 수 있다. box plot은 데이터의 분산 정도를 최댓값, 제 3사분위수, 중간값, 제 1사분위수, 최솟값으로 요약하여 나타낸다. iqr은 제 3사분위수에서 제 1사분위수를 뺀 값이며, iqr rule은 제 1사분위수, 제 3사분위수에서 iqr의 1.5배 이상 벗어난 값을 이상치로 판단하는 방법이다(Kim, 2017).

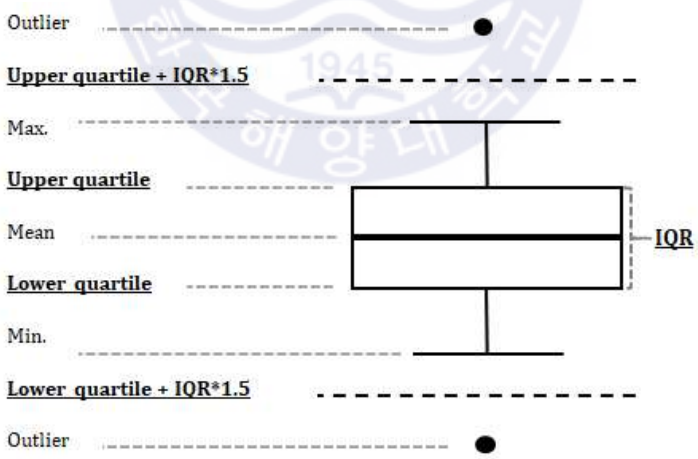


Fig. 11 Method of checking outlier using iqr rule

② cook’s distnace

회귀 분석에서는 레버리지와 잔차의 크기가 큰 데이터를 이상치라고 부르며, cook’s distance는 레버리지와 잔차를 동시에 보기위한 기준이 된다. 레버리지란, 실제 결과 값 y가 예측 값 $\hat{y}$에 미치는 영향을 나타낸 값이다. 영향도 행렬 H에 대해 $\hat{y} = Hy$라는 식을 만족하며, 레버리지는 수학적으로 영향도 행렬의 대각성분인 h_{ii}으로 정의된다. 잔차란, 표본의 회귀식으로부터 추정된 값과 실제 값의 차이를 의미한다. 잔차의 크기는 독립 변수의 영향을 받기 때문에 레버리지와 잔차의 표준 편차로 나누어 동일한 표준 편차를 가지도록 스케일링한 표준화된 잔차 r_i를 사용해야 한다.

cook’s distance를 식 (3)과 같이 표현되며 레버리지가 커지거나 잔차의 크기가 커지면 cook’s distance 값이 커진다.

$$D_i = \frac{ri^2}{RSS}[\frac{h_{ii}}{(1-h_{ii})^2}] \tag{3}$$

cook’s distance가 식 (4)와 같은 기준 값보다 클 때 이상치로 판단하게 되며, N은 데이터 개수, K는 레버리지 합을 의미한다.

$$D_i > \frac{4}{N-K-1} \tag{4}$$

데이터 분석 시 이상치를 포함하면 분석결과의 정확성이 떨어지게 되므로, 발견된 이상치를 적절한 방법으로 처리해 주어야 한다. 이상치를 처리하는 방법은 가장 간단한 방법으로는 이상치를 삭제하고 분석하는 것이다. 또는, 데이터 값을 로그변환 함으로써 극단적인 값들의 효과를 감소시키는 방법과 이상치를 평균이나 중앙값으로 대체하는 방법이 있다. 이상치가 측정 또는 입력 오류에 의해서 발생한 경우에는 해당 관측치를 제거하면 되지만, 데이터가 절대적으로 적은 경우에는 제거하는 방법으로 이상치를 처리하면 관측치가 적어지는 문제가 발생하기도 한다. 자연 발생한 이상치를 단순 삭제나 대체의 방법으로 처리하여 모델을 만든 경우 설명하고자하는 현상을 잘 성명하지 못할 수도 있기 때문에, 이상치를 바로 삭제하기 보다는 좀 더 찬찬히 이상치에 대해 파악하는 것이 중요하다.

2.2.2.4 원-핫 인코딩

데이터에는 수치형 데이터뿐만 아니라 문자형 데이터인 범주형 데이터 또한 존재한다. 그러나 범주형 데이터들은 일반적인 수치형 데이터들과는 다르게 컴퓨터가 바로 인식할 수 없기 때문에 컴퓨터가 인식할 수 있는 형태 즉, 문자를 숫자로 변환하는 과정이 필요하다.

원-핫 인코딩은 문자를 숫자로 바꾸는 여러 가지 기법 중에서 단어를 표현하는 가장 기본적인 방법이다. 단어 집합의 크기를 벡터의 차원으로 변환하고 표현하고 싶은 단어의 인덱스에 1, 나머지에는 0을 부여하는 단어의 벡터 표현 방식이다. 이 방법은 지금까지 좋은 성능을 내고 지금까지도 많은 사람들이 사용하지만, 컴퓨터가 단어의 의미 또는 개념 차이를 전혀 담지 못한다는 단점이 있다.

2.2.3 모델 학습

모델 학습 단계에서 모델이란 새로운 입력 데이터를 받았을 때 예측 값을 계산하는 방법을 의미하며, 일반적으로 예측 값을 계산하는 알고리즘이 예측 모델이 된다. 세상에는 수많은 종류의 알고리즘이 있고 주어진 문제와 데이터에 맞는 적절한 알고리즘을 선택하는 것은 데이터 분석가의 몫이다.

모델 학습 단계에서 적절한 알고리즘을 선택한다는 것은 첫째, 말 그대로 예측 값을 계산하기 위한 알고리즘을 선택하는 것이다. 수많은 종류의 알고리즘 중에서 여러 가지 알고리즘을 선택하여 예측을 시도해 보고 데이터에 가장 적합한 알고리즘을 선택해야 한다. 둘째, 알고리즘이 사용할 속성들을 선택하는 것이다. 때때로 의미 없는 속성이 학습에 사용될 때 알고리즘의 성능이 더 떨어지는 경우가 있다. 따라서 중요한 속성들을 골라내는 일도 알고리즘 선택 과정에서 필요한 일이다. 셋째, 알고리즘에는 일종의 알고리즘의 성능을 조절하는 하이퍼파라미터(hyper-parameter)를 선택하는 것이다. 같은 알고리즘을 사용하더라도 하이퍼파라미터에 따라서 성능이 천차만별이기 때문에 이를 선택하는 것 또한 중요한 과정이라고 할 수 있다.

2.2.4 학습 모델 성능 확인

학습 모델 성능 확인 단계에서는 만들어진 학습 모델의 성능을 확인하는 단계이다. 해당 단계에서 반드시 지켜야 할 점은 평가용 데이터 셋은 모델 선택과 모델 학습과정에서 쓰이지 않아야 한다는 점이다. 즉, 프로젝트를 시작하기 전에 학습용 데이터 셋과 평가용 데이터 셋을 나누어놓고, 평가용 데이터 셋은 모델 학습 단계가 끝나기 전까지는 보지 말아야 한다는 것이다. 이렇게 하는 이유는 학습 모델 성능 확인 과정의 목적이 모델이 새로운 데이터에 대해 얼마나 일반화 가능한지 측정하는 것이기 때문이다.

본 연구에서는 여러 가지 평가 지표를 활용하여 학습 모델의 성능을 확인하고자 한다. 학습 모델에서 산출된 예측 값과 실제 데이터의 실적 값 사이의 오차 및 정확도를 정량적 지표로 산출하기 위해 MAE(mean absolute error), MAPE(mean squared percentage error), RMSE(root mean square error), RMSLE(root mean squared logarithmic error)의 4가지 지표를 활용하고자 하며 자세한 내용은 Fig. 12와 같다.

	validation criteria	equation
오차기준	MAE (mean absolute error)	$\lvert y_i - \hat{y}_i \rvert$
오차기준	MAPE (mean squared percentage error)	$\frac{100}{n}\sum_{i=1}^{n}\lvert (y_i - \hat{y}_i)/y_i \rvert$
잔차기준	RMSE (root mean square error)	$\sqrt{\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]/n}$
잔차기준	RMSLE (root mean squared logarithmic error)	$\sqrt{\sum_{i=1}^{n}(\log(\hat{y}_i + 1) - \log(y_i + 1))^2/n}$

(y_i : 실적 값, $\hat{y}_i$: 예측 값)

Fig. 12 Criteria for learning model performance validation

MAE는 평균 절대 오차로 예측 값과 실적 값의 차이를 절댓값으로 나타낸 것을 의미하며, MAPE는 평균 절대 오차 비율로 예측 값과 실적 값의 차이를 백분율로 환산한 값이다. RMSE는 평균 제곱근 오차로 엄밀히 따지자면 잔차를 제곱해서 평균한 값의 제곱근을 의미하며, 보통 정밀도라고 표현되는 값이다. 마지막으로 RMSLE는 평균 제곱근 로그 오차 비율로 잔차의 평균에 로그를 씌운 값이다. 예측의 오차는 이상치가 큰 부분에서 발생하기도 하지만 작은 부분으로부터 발생하기도 하는데, RMSLE는 이상치가 과대평가된 항목 보다는 과소평가된 항목에 페널티를 주어 오차를 계산한다. 이 때, MAE와 MAPE는 모집단의 회귀 식으로부터 예측한 값과 실제 값의 차이인 오차를 기준으로 하며, RMSE와 RMSLE는 표본 집단의 회귀 식으로부터 예측한 값과 실제 값의 차이인 잔차를 기준으로 한다.

본 연구에서는 오차를 기준으로 오차가 갖는 영향력의 정도를 가장 직관적으로 확인할 수 있는 값을 MAPE라고 판단하였기 때문에 학습 모델 성능 확인 단계에서 MAPE를 가장 우선적으로 고려한다.

제 3 장 조선 생산 리드타임 예측을 위한 데이터 분석

조선 생산 리드타임 예측을 위한 데이터 분석을 수행하기에 앞서 선박의 건조 과정을 간략하게 설명하고자 한다. 선박은 일반 기계, 자동차, 전기 제품 등과 달리 선주의 주문에 의해 생산하게 되는 주문생산 방식을 취한다. 선주와 조선소간의 건조계약이 체결되면 조선소는 건조계획을 수립하는 한편 선주가 요구하는 사양에 맞춰 설계를 하게 된다. 다음으로 생산 제품별로 강재를 적치하는 강재적치, 선체구조를 구성하고 있는 부재를 제작하는 공정인 가공 공정을 수행한다. 선체구조 부재는 평면판, 곡면판, 직선형강 및 곡선형강 등으로 구성되며 이 부재들은 강판재와 형강재의 절단작업, 굽힘 작업을 통해 제작된다. 가공 공정을 통해 제작된 선체의 부재는 조립장으로 이송되어 선체 내부 구조물에 보강재를 붙이는 소조립 과정을 거쳐 선체 외판재에 늑골을 붙이는 중조립 과정으로 이어지고, 다시 블록을 완성하는 대조립 과정을 거친다. 또한 선박의 일부분인 블록에 파이프나 배선 등의 의장작업을 하게 되고, 배가 녹슬지 않도록 친환경적 제품으로 페인트 작업을 하는 도장 단계를 거친 뒤 완성된 블록을 도크로 옮겨 탑재하여 선박의 모양을 갖추게 된다. 해당 단계에서는 설계도에 제시된 선체의 형상이 완성되어야 하기 때문에 선체의 치수가 정확한지 점검해야 한다. 다음으로 도크에 물을 채워 완성된 선박을 바다로 띄우게 되고, 진수된 선박을 안벽에서 선실의 인테리어 및 각종 장비를 설치하고 테스트를 하며 작업을 마무리한다. 마지막으로 해상에서 선박의 성능을 최종적으로 테스트한 뒤 선주가 직접 방문하여 완성된 선박의 이름을 부여하는 행사를 함으로써 선박 건조 공정을 마무리한다.

선박의 건조 과정을 그림으로 나타내면 Fig. 13와 같으며, 본 연구에서는 강재 절단, 조립, 의장 공정에 해당하는 데이터를 수집하여 분석을 수행하였다.

Fig. 13 Shipbuilding process

3.1 해양플랜트 배관재 공급망 데이터 분석

본 연구에서는 해양플랜트 공급망의 생산 리드타임을 개선하기 위해 의장 공정에 해당하는 해양플랜트의 배관재 공급망 데이터를 분석하였다. 조선소에서의 선박 및 해양플랜트 건조는 설계부터 생산까지 복잡한 공정으로 이루어져 있을 뿐만 아니라 선체에 조립되는 다양한 종류의 의장품들 또한 복잡한 공급망을 통해 관리되고 있다. 특히 해양플랜트 의장 공정의 대부분을 차지하고 있는 배관재는 적절한 조달관리가 어렵고 수작업에 따른 관리의 한계로 납기 지연에 따른 문제점이 발생하는 경우가 있다.

배관재 공급망은 크게 6개의 공정으로 제작 공정부터 설치 공정까지 공정절점별로 리드타임이 관리되고 있다. W/O 발행일을 시작으로 제작(making), 도장(painting), 사외적치(out stock), 사내적치(in stock), 설치대기(standing install), 설치(install) 순으로 공정이 진행된다. 배관 공정의 일반적인 흐름도는 Fig. 14와 같다.

Fig. 14 Spool procurement process

제작 공정은 배관재를 제작하는 공정기간으로 제작지시가 발행된 시점부터 제작이 완료되기까지의 기간을 의미한다. 제작순서는 보통 발주된 순서대로 진행하지만 긴급 물량이나 품질검사 결과에 따라 완료날짜가 변경될 수 있다. 도장 공정은 제작이 완료된 시점부터 도장이 완료되는 시점까지를 의미한다. 이 단계에서는 배관재에 따라 수행여부가 달라지기도 하는데, 후행도장 대상 품목이나 긴급 입고인 경우에는 해당 과정을 거치지 않고 제작 후 바로 야드 내로 입고되는 경우가 존재한다. 사외적치 공정은 도장이 완료되고 난 뒤 야드에 입고되기 전에 잠깐 사외 적치장에 적치되어 있는 기간으로 도장이 없는 배관재는 해당 적치장을 거치지 않고 바로 야드로 입고된다. 사내적치 공정은 야드에 입고된 후 야드에 입고된 후 배관재가 작업장

주위로 적치되기까지의 기간을 의미한다. 이 단계의 경우 해당 적치장에 각 호선에 설치되는 하나의 배관만 적치되는 것이 아니라 여러 호선의 배관재들을 동시에 관리하기 때문에 이를 찾는데 소요되는 시간이 많을 뿐 아니라 분실 또한 빈번히 일어나고 있어 이러한 요소들이 특이점으로 작용할 수 있는 가능성을 가지고 있다. 설치대기 공정은 배관재가 실제 설치될 장소에 적치되기까지의 기간을 의미하며, 설치 공정은 배관재가 설치 장소에 적치된 후 설치가 최종 완료되는 시점까지를 의미한다. 본 논문의 선행연구인 함동균(2016)에서는 전체 6개의 공정에 따른 리드타임을 예측하였지만 사외적치 이후의 공정에서는 다양한 외적요인 및 데이터의 결함으로 인해 예측도가 현저히 떨어진 것으로 판단하여 본 논문에서는 제작과 도장 공정의 리드타임만을 대상으로 하였다.

본 연구에서는 해양플랜트 공급망의 생산 리드타임을 개선하기 위해 해양플랜트의 배관재 공급망 데이터를 분석하고자 한다. 이를 위해 S사의 해양플랜트 배관재 공급망 데이터를 수집하였고, 각 공정에 따른 생산 리드타임을 예측하기 위해 다양한 기계학습, 심층학습, 앙상블학습 알고리즘을 적용하였다. 데이터 분석을 위한 도구로는 Python에서 제공하는 다양한 라이브러리를 활용하였으며, 데이터 분석 과정에 따라 데이터 전처리를 수행한 후 각 알고리즘에 따른 생산 리드타임 예측 모델을 생성하고 평가지표를 활용하여 생성된 예측 모델의 성능을 확인하였다.

데이터 분석을 수행하기에 앞서 실제 조선소로부터 해양플랜트의 배관재 공급망 데이터를 수집한 결과, 비교적 최근에 작업된 하나의 호선에 설치된 배관재 데이터 32,029개를 수집할 수 있었고 1차적으로 선별한 데이터는 Table 1과 같다. 해당 데이터는 5개의 연속형 변수와 8개의 범주형 변수까지 총 13개의 독립변수와 제작, 도장 공정의 리드타임에 해당하는 종속변수로 이루어져 있다.

Table 1 Spool procurement process data

Data	Contents	
Collection Data	Raw Data (32,039 rows)	
Input Data	Continuous Variable	DIA
		Length
		Weight
		Member Count
		Joint Count
	Categorical Variable	Emergency
		Apply Lead Time
		STG
		Service
		Sch
		Material
		Making Co.
		After2 Co.
Output Data	Lead time (days)	

가장 먼저 상관분석과 분산분석을 수행하여 종속변수인 리드타임과 유의미한 관계를 갖는 독립변수를 선별하였다. 독립변수인 5개의 연속형 변수와 종속변수에 해당하는 제작, 도장 공정의 리드타임 사이의 상관관계를 분석하기 위해 상관분석을 수행하였다. 상관분석을 수행한 결과는 Fig. 15와 같으며 DIA와 Weight, Member Count와 Joint Count의 상관계수의 경우 각각 0.76, 0.85로 기준 값인 0.65보다 큰 값을 갖지만, 현장 작업자와의 인터뷰를 통해 해당 변수들이 종속변수인 리드타임을 결정하기 위한 중요한 요소이므로 분석에 포함되었으면 한다는 의견을 반영하여 변수에서 제외하지 않았다. 따라서 최종적으로 분석에 활용될 독립변수에 해당하는 연속형 변수는 DIA, Length, Weight, Member Count, Joint Count이다.

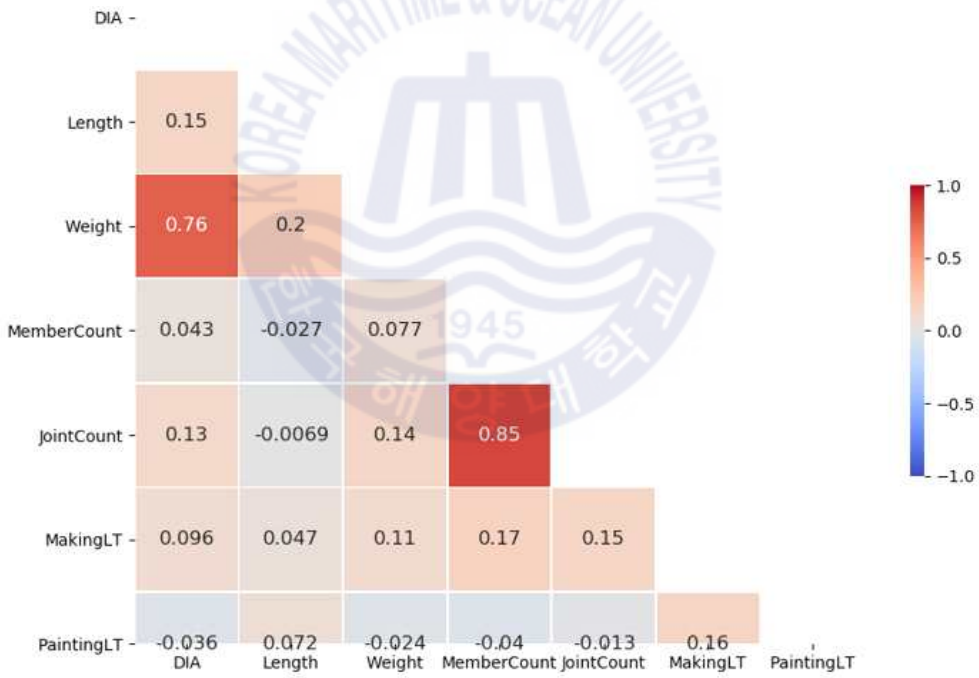


Fig. 15 Result of correlation analysis (a)

범주형 변수와 종속변수인 연속형 변수 사이의 관계는 분산분석을 통해 확인하였다. 분산분석은 F-value와 P-value로 판단이 가능하며, 독립변수인 8개의 범주형 변수와 종속변수에 해당하는 제작, 도장 공정의 리드타임 사이의 분산분석을 수행하였다. 분산분석을 수행한 결과는 Table 2와 같으며 F-value 및 P-value 값이 유의수준인 0.05에 미치지 못한 변수는 존재하지 않았기 때문에 1차적으로 선별한 8개의 범주형 변수를 모두 분석에 활용하고자 한다. 따라서 최종적으로 분석에 활용될 독립변수에 해당하는 범주형 변수는 Emergency, Apply Lead Time, STG, Service, Sch, Material, Making Co., After2 Co.이다.

Table 2 Result of ANOVA (a)

	Sum Sq	Df	F value	PR(>F)
Emergency	1.435e+05	1	1330.855	1.110e-284
Apply Lead Time	7.655e+02	2	3.548	2.879e-02
STG	1.560e+04	7	20.669	6.303e-28
Service	2.348e+04	47	4.631	1.110e-23
Sch	1.464e+03	2	6.787	1.129e-03
Material	1.508e+05	5	279.741	5.778e-293
Making Co.	6.743e+04	6	104.191	2.649e-130
After2 Co.	4.254e+04	6	65.734	1.726e-81

상관분석과 분산분석의 결과를 적용하여 최종적으로 정의된 독립변수는 5개의 연속형 변수와 8개의 범주형 변수까지 총 13개이며 이를 활용하여 종속변수인 제작, 도장 공정의 리드타임을 예측하고자 한다.

해양플랜트 배관재 공급망 데이터 분석을 위해 최종적으로 선별된 독립변수와 종속변수로만 구성된 데이터를 정리한 뒤, 데이터 내에 존재하는 결측값을 단순 제거를 통해 처리해 주었다. 또한 종속변수인 제작, 도장 리드타임의 이상치를 iqr rule, cook's distance를 통해 확인하여 단순 제거를 통해 Fig. 16, 17과 같이 처리해 주었다.

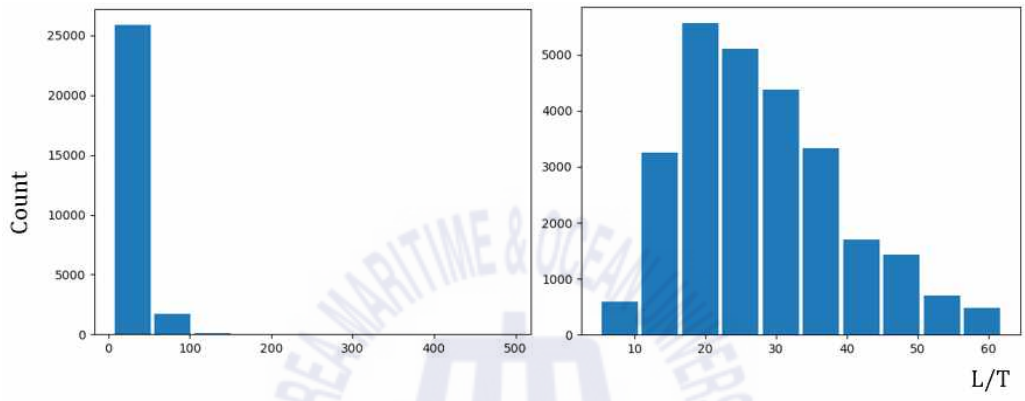


Fig. 16 Processing of the outlier of the spool making process

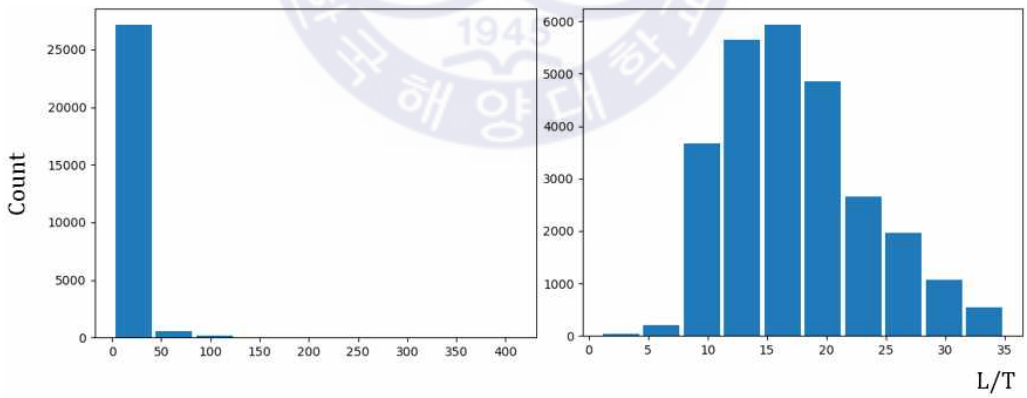


Fig. 17 Processing of the outlier of the spool painting process

마지막으로 범주형 변수를 컴퓨터가 인식할 수 있는 형태, 즉 문자를 숫자로 변환하는 one-hot encoding 단계를 거치면서 데이터 전처리를 마무리 하였다.

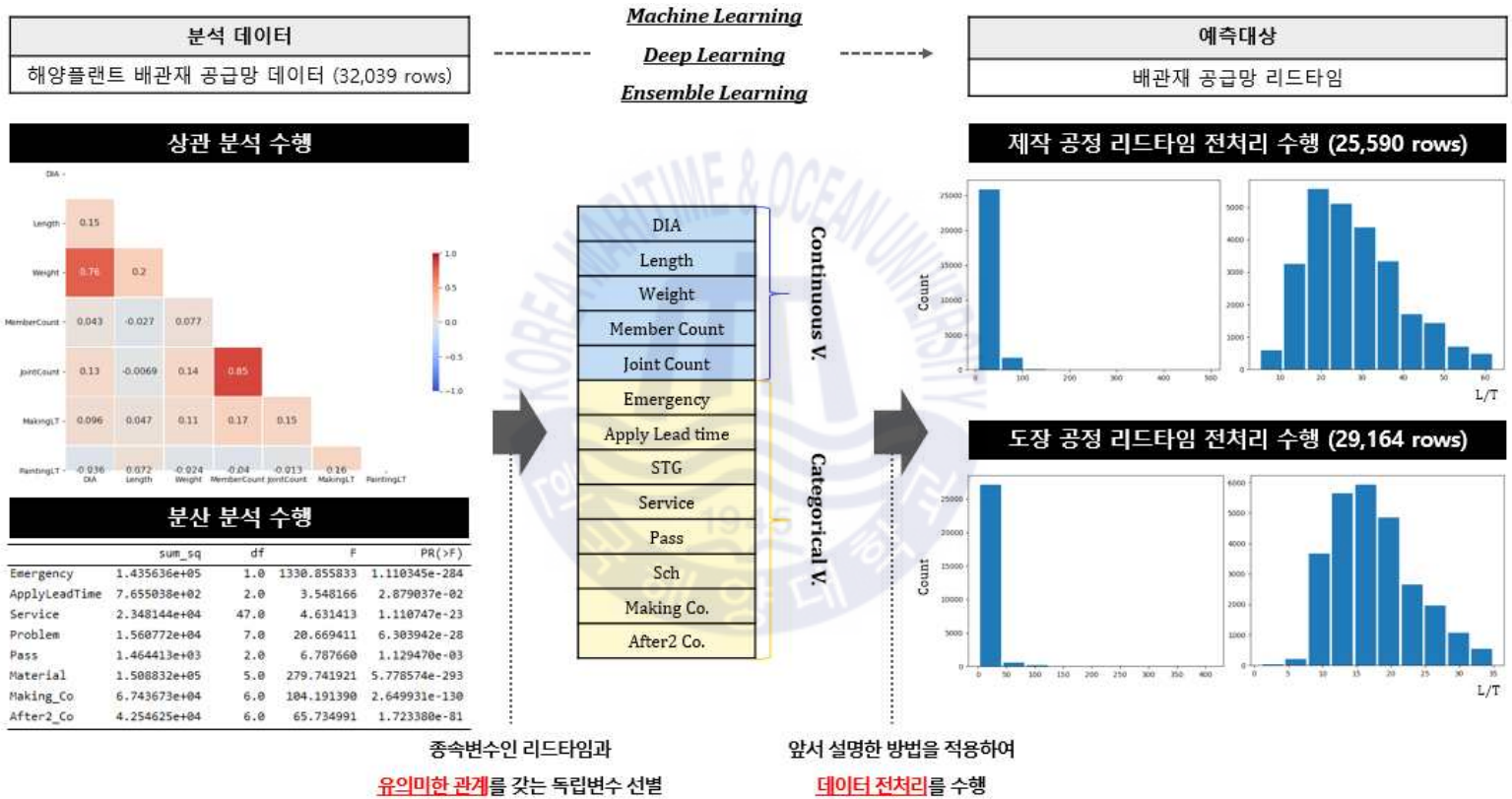


Fig. 18 Spool procurement process data analysis process

3.2 조선소 블록 조립 공정 데이터 분석

본 연구에서는 조선소 블록 조립 공정의 생산 리드타임을 개선하기 위해 조립 공정에 해당하는 조선소 블록 조립 공정 데이터를 분석하였다. 선박은 블록 건조 공법으로 생산 되는데, 이는 거대한 선박을 작업장 내에서 조립할 수 있는 적절한 크기의 블록으로 나누고, 이들 블록들을 작업장 내에서 조립한 후, 의장, 도장 작업 등을 거쳐 도크에서 최종 탑재하여 선박을 건조하는 방식이다. 이때 조립 작업장 내에서 조립되는 단위 블록들을 조립 블록이라고 한다. 조립 블록은 작업장 내에서 작업할 수 있는 수준의 크기로 분할되지만 가로, 세로의 길이가 수 미터에 이르고, 무게가 수십 톤에 달하며, 많은 수의 조립품들이 여러 조립 작업 단계를 거쳐서 완성된다. 또한 하나의 조립 블록은 선박의 일부분이므로 구조적 특성상 좌/우현의 유사블록이 존재할 뿐 동일한 블록이 없고 이를 이루는 조립품의 구성도 모두 다른 특징을 가진다. 따라서 생산 계획자는 각 조립 블록에 대해서 매번 작업 단계를 정의하고 생산계획을 수립해야 하는 어려움이 있다. 또한 조선 산업의 특성상 작업장이 넓고 협력업체 등 많은 작업 조직과 작업자에 의해서 생산이 이루어지기 때문에 실제 생산 공정을 모두 이해하고 파악하는 것이 힘들며 이는 정확한 생산 계획과 관리를 어렵게 하는 요인이 된다.

본 연구에서는 조선소 블록 조립 공정의 생산 리드타임을 개선하기 위해 조선소 블록 조립 공정의 데이터를 분석하고자 한다. 이를 위해 D사의 조선소 블록 조립 공정 데이터를 수집하였고, 생산 리드타임을 예측하기 위해 다양한 기계학습, 심층학습, 앙상블학습 알고리즘을 적용하였다. 데이터 분석을 위한 도구로는 Python에서 제공하는 다양한 라이브러리를 활용하였으며, 데이터 분석 과정에 따라 데이터 전처리를 수행한 후 각 알고리즘에 따른 생산 리드타임 예측 모델을 생성하고 여러 가지 평가지표를 활용하여 생성된 예측 모델의 성능을 확인하였다.

데이터 분석을 수행하기에 앞서 실제 조선소로부터 블록 조립 공정의 데이터를 수집한 결과, 약 4년간 저장된 블록 조립 공정 실적 데이터 11,243개를 수집할 수 있었고 1차적으로 선별한 데이터는 Table 3와 같다. 해당 데이터는 7개의 연속형 변수와 9개의 범주형 변수까지 총 16개의 독립변수와 블록 조립 공정의 리드타임에 해당하는 종속변수로 이루어져 있다.

Table 3 Block assembly process data

Data	Contents	
Collection Data	Raw Data (11,243 rows)	
Input Data	Continuous Variable	Weight
		Length
		Breadth
		Height
		Plan Lead Time
		Rain
		Snow
	Categorical Variable	Team
		Process
		Ship Type
		Project
		Block
		Assembly
		PCG
		PCG Name
		Business
Output Data	Lead time (days)	

가장 먼저 상관분석과 분산분석을 수행하여 종속변수인 리드타임과 유의미한 관계를 갖는 독립변수를 선별하였다. 독립변수인 7개의 연속형 변수와 종속변수에 해당하는 블록 조립 공정의 리드타임 사이의 상관관계를 분석하기 위해 상관분석을 수행하였다. 상관분석을 수행한 결과는 Fig. 19와 같으며 상관 계수의 기준 값인 0.65보다 큰 값을 갖는 변수는 존재하지 않았기 때문에 1차적으로 선별한 7개의 연속형 변수를 모두 분석에 활용하고자 한다. 따라서 최종적으로 분석에 활용될 독립변수에 해당하는 연속형 변수는 Weight, Length, Breath, Height, Plan Lead Time, Rain, Snow이다.

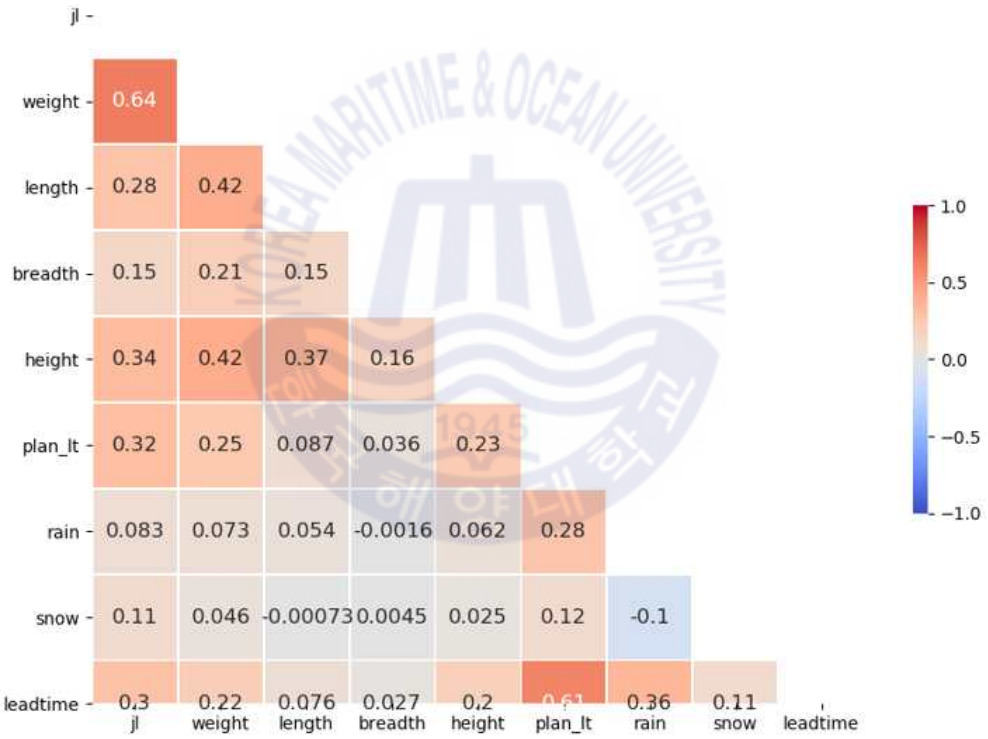


Fig. 19 Result of correlation analysis (b)

범주형 변수와 종속변수인 연속형 변수 사이의 관계는 분산분석을 통해 확인하였다. 분산분석은 F-value와 P-value로 판단이 가능하며, 독립변수인 9개의 범주형 변수와 종속변수에 해당하는 블록 조립 공정의 리드타임 사이의 분산분석을 수행하였다. 분산분석을 수행한 결과는 Table 4와 같으며 F-value 및 P-value 값이 유의수준인 0.05에 미치지 못한 변수는 존재하지 않았기 때문에 1차적으로 선별한 9개의 범주형 변수를 모두 분석에 활용하고자 한다. 따라서 최종적으로 분석에 활용될 독립변수에 해당하는 범주형 변수는 Team, Process, Ship Type, Project, Block, Assembly, PCG, PCG Name, Business이다.

Table 4 Result of ANOVA (b)

	Sum Sq	Df	F value	PR(>F)
Team	48.030	1	7.674	5.610e-03
Process	255.926	1	10.893	1.674e-10
Ship Type	1140.250	6	30.365	5.416e-36
Project	8839.163	43	30.987	1.370e-235
Block	30143.324	304	15.843	0.000e+00
Assembly	29682.421	150	31.618	0.000e+00
PCG	14342.765	22	104.170	0.000e+00
PCG Name	14401.772	23	100.051	0.000e+00
Business	7058.675	44	25.633	2.219e-196

상관분석과 분산분석의 결과를 적용하여 최종적으로 정의된 독립변수는 7개의 연속형 변수와 9개의 범주형 변수까지 총 16개이며 이를 활용하여 종속변수인 블록 조립 공정의 리드타임을 예측하고자 한다.

조선소 블록 조립 공정 데이터 분석을 위해 최종적으로 선별된 독립변수와 종속변수로만 구성된 데이터를 정리한 뒤, 데이터 내에 존재하는 결측값을 확인하여 단순 제거를 통해 처리해 주었다. 또한 종속변수인 제작, 도장 리드타임의 이상치를 iqr rule, cook's distance를 통해 확인하여 단순 제거를 통해 Fig. 20과 같이 처리해 주었다.

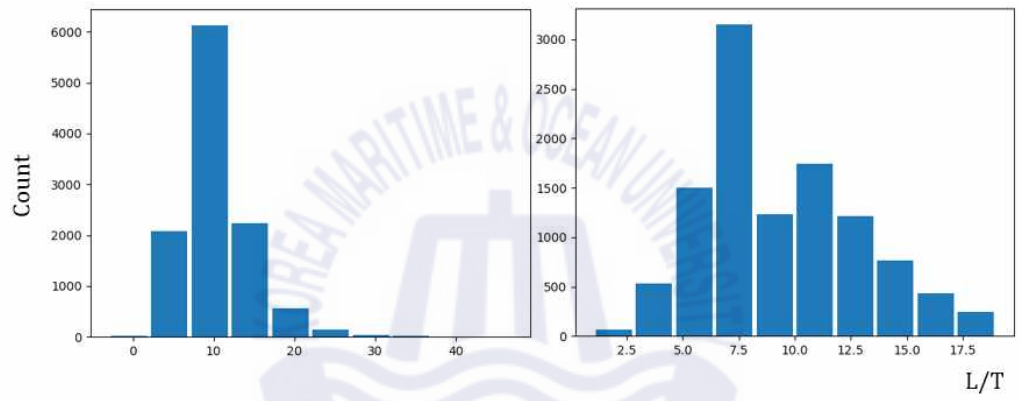
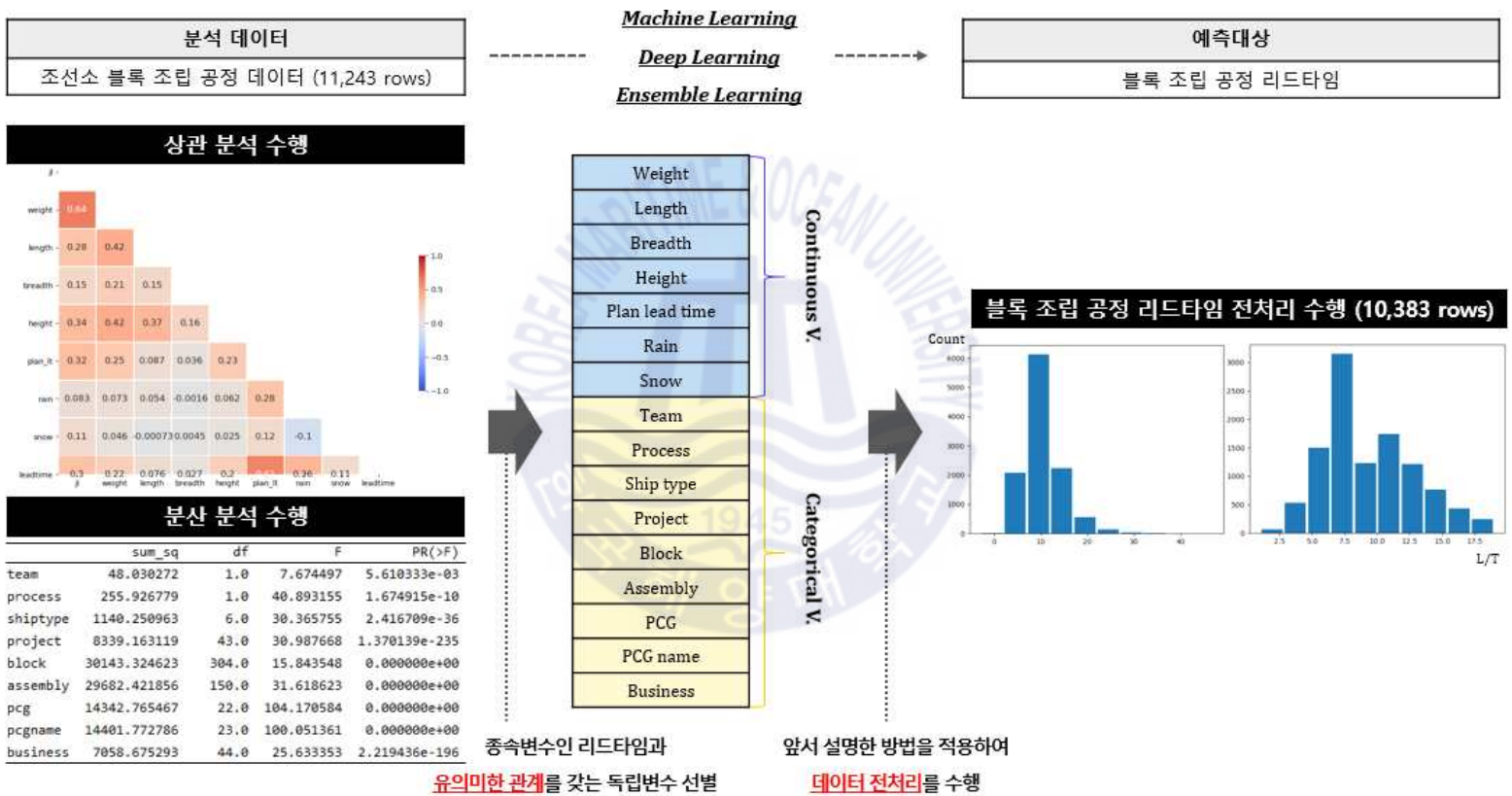


Fig. 20 Processing of the outlier of the block assembly process

마지막으로 범주형 변수를 컴퓨터가 인식할 수 있는 형태, 즉 문자를 숫자로 변환하는 one-hot encoding 단계를 거치면서 데이터 전처리를 마무리 하였다.



	sum_sq	df	F	PR(>F)
team	48.030272	1.0	7.674497	5.610333e-03
process	255.926779	1.0	40.893155	1.674915e-10
shiptype	1140.250963	6.0	30.365755	2.416709e-36
project	8339.163119	43.0	30.987668	1.370139e-235
block	30143.324623	304.0	15.843548	0.000000e+00
assembly	29682.421856	150.0	31.618623	0.000000e+00
pcg	14342.765467	22.0	104.170584	0.000000e+00
pcgname	14401.772786	23.0	100.051361	0.000000e+00
business	7058.675293	44.0	25.633353	2.219436e-196

Fig. 21 Block assembly process data analysis process

3.3 조선소 블록 절단 공정 데이터 분석

본 연구에서는 조선소 블록 절단 공정의 생산 리드타임을 개선하기 위해 강재 절단 공정에 해당하는 조선소 블록 절단 공정 데이터를 분석하였다. 일반적으로 선박을 건조하는데 있어서, 강재는 원재료비의 대략 30-35%를 차지할 정도로 많은 비중을 차지하고 있는 중요한 자재이며 물성의 특성상 하중과 부피가 커서 효율적인 관리가 어려워 많은 양의 강재를 관리하기 위한 시간과 비용을 필요로 하고 있다. 조선소에서 강재는 선박 건조에 있어서 강재의 적기 수급 및 투입의 관점에서 매우 중요한 관리 역량을 필요로 하고 있다. 예를 들어 해당 블록에 포함되는 하나의 부재라도 적시에 작업 투입이 되지 못하면 해당 블록에 대한 생산 공정 및 일정에 영향을 미치게 되며, 이로 인하여 작업 및 공기 지연이 발생하게 되어 이후 생산 일정에 많은 문제점을 발생시킨다. 그러므로 요구되어지는 강재가 생산일정에 정확하게 투입되는 것은 매우 중요하다. 그리고 강재는 일반 자재와 달리 자재 수급 소요기간이 길고 관리가 부실할 경우 물성의 특성상 무게와 부피로 인하여 적치, 선별하는 작업에 많은 어려움이 있어 업무처리에 많은 소요 시간을 필요로 한다. 그렇기 때문에 조선소의 강재관리 업무의 특성을 반영한 체계적인 관리가 되지 않을 경우, 정확한 관리가 불가능하다.

본 연구에서는 조선소 블록 절단 공정의 생산 리드타임을 개선하기 위해 조선소 블록 절단 공정의 데이터를 분석하고자 한다. 이를 위해 S’ 사의 조선소 블록 절단 공정 데이터를 수집하였고, 생산 리드타임을 예측하기 위해 다양한 기계학습, 심층학습, 앙상블학습 알고리즘을 적용하였다. 데이터 분석을 위한 도구로는 Python에서 제공하는 다양한 라이브러리를 활용하였으며, 데이터 분석 과정에 따라 데이터 전처리를 수행한 후 각 알고리즘에 따른 생산 리드타임 예측 모델을 생성하고 여러 가지 평가지표를 활용하여 생성된 예측 모델의 성능을 확인하였다.

데이터 분석을 수행하기에 앞서 실제 조선소로부터 블록 절단 공정의 데이터를 수집한 결과, 약 6년간 저장된 블록 절단 공정 실적 데이터 63,989개를 수집할 수 있었고 1차적으로 선별한 데이터는 Table 5과 같다. 해당 데이터는 3개의 연속형 변수와 4개의 범주형 변수까지 총 7개의 독립변수와 블록 절단 공정의 리드타임에 해당하는 종속변수로 이루어져 있다.

Table 5 Block cutting process data

Data	Contents	
Collection Data	Raw Data (63,989 rows)	
Input Data	Continuous Variable	Weight
		Rain
		Plan Lead Time
	Categorical Variable	Block Group
		Block Position
		Ship Type
		Business
Output Data	Lead time (days)	

가장 먼저 상관분석과 분산분석을 수행하여 종속변수인 리드타임과 유의미한 관계를 갖는 독립변수를 선별하였다. 독립변수인 3개의 연속형 변수와 종속변수에 해당하는 블록 절단 공정의 리드타임 사이의 상관관계를 분석하기 위해 상관분석을 수행하였다. 상관분석을 수행한 결과는 Fig. 22와 같으며 Rain과 Lead Time 사이의 상관계수가 0.97로 기준 값인 0.65보다 큰 값을 갖는 것으로 확인하여 Rain을 독립변수에서 제외하고자 한다. 따라서 최종적으로 분석에 활용될 독립변수에 해당하는 연속형 변수는 Weight, Plan Lead Time이다.

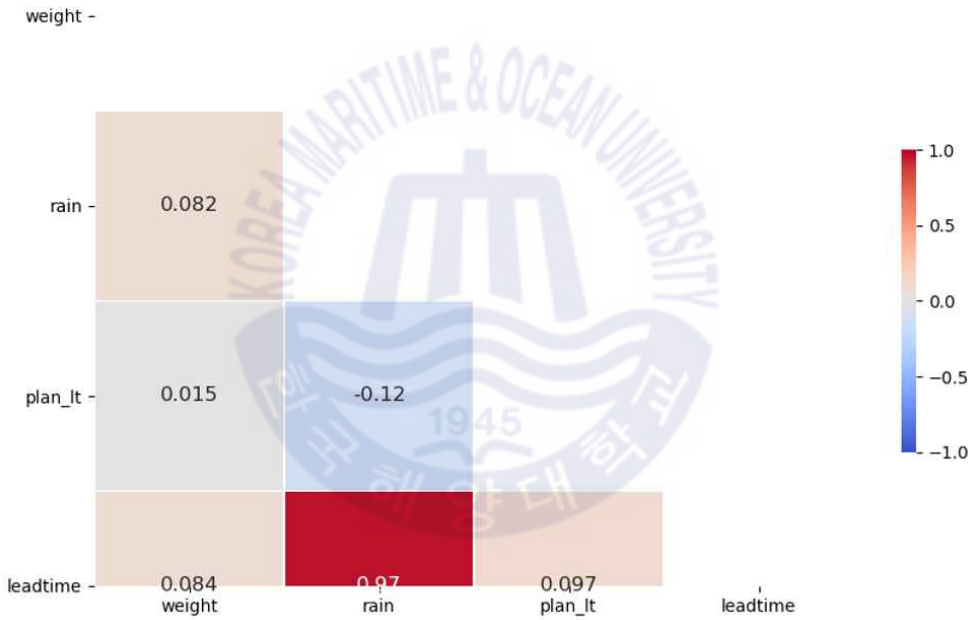


Fig. 22 Result of correlation analysis (c)

범주형 변수와 종속변수인 연속형 변수 사이의 관계는 분산분석을 통해 확인하였다. 분산분석은 F-value와 P-value로 판단이 가능하며, 독립변수인 4개의 범주형 변수와 종속변수에 해당하는 블록 절단 공정의 리드타임 사이의 분산분석을 수행하였다. 분산분석을 수행한 결과는 Table 6와 같으며 F-value 및 P-value 값이 유의수준인 0.05에 미치지 못한 변수는 존재하지 않았기 때문에 1차적으로 선별한 9개의 범주형 변수를 모두 분석에 활용하고자 한다. 따라서 최종적으로 분석에 활용될 독립변수에 해당하는 범주형 변수는 Block Group, Block Position, Ship Type, Business이다.

Table 6 Result of ANOVA (c)

	Sum Sq	Df	F value	PR(>F)
Block Group	7.002e+05	14	103.537	1.013e-297
Block Position	6.544e+04	2	67.733	4.120e-30
Ship Type	4.641e+06	13	739.041	0.000e+00
Business	1.516e+06	61	51.457	0.000e+00

상관분석과 분산분석의 결과를 적용하여 최종적으로 정의된 독립변수는 2개의 연속형 변수와 4개의 범주형 변수까지 총 6개이며 이를 활용하여 종속변수인 블록 절단 공정의 리드타임을 예측하고자 한다.

조선소 블록 절단 공정 데이터 분석을 위해 최종적으로 선별된 독립변수와 종속변수로만 구성된 데이터를 정리한 뒤, 데이터 내에 존재하는 결측값을 확인하여 단순 제거를 통해 처리해 주었다. 또한 종속변수인 제작, 도장 리드타임의 이상치를 iqr rule, cook's distance를 통해 확인하여 단순 제거를 통해 Fig. 23과 같이 처리해 주었다.

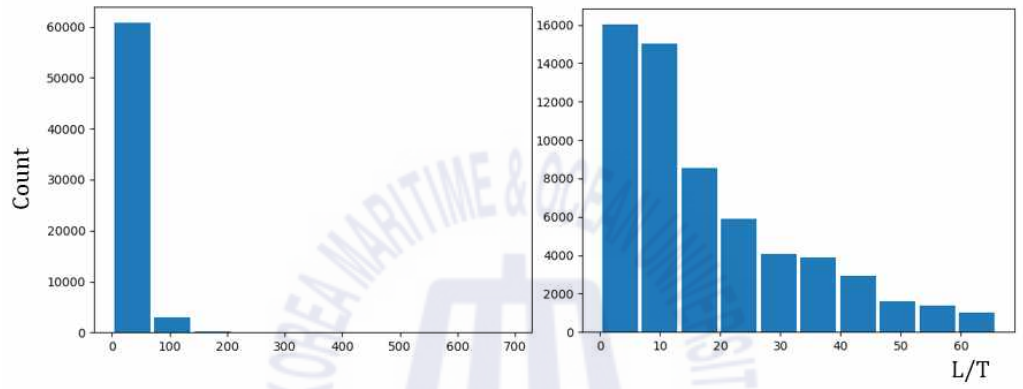


Fig. 23 Processing of the outlier of the block cutting process

마지막으로 범주형 변수를 컴퓨터가 인식할 수 있는 형태, 즉 문자를 숫자로 변환하는 one-hot encoding 단계를 거치면서 데이터 전처리를 마무리 하였다.

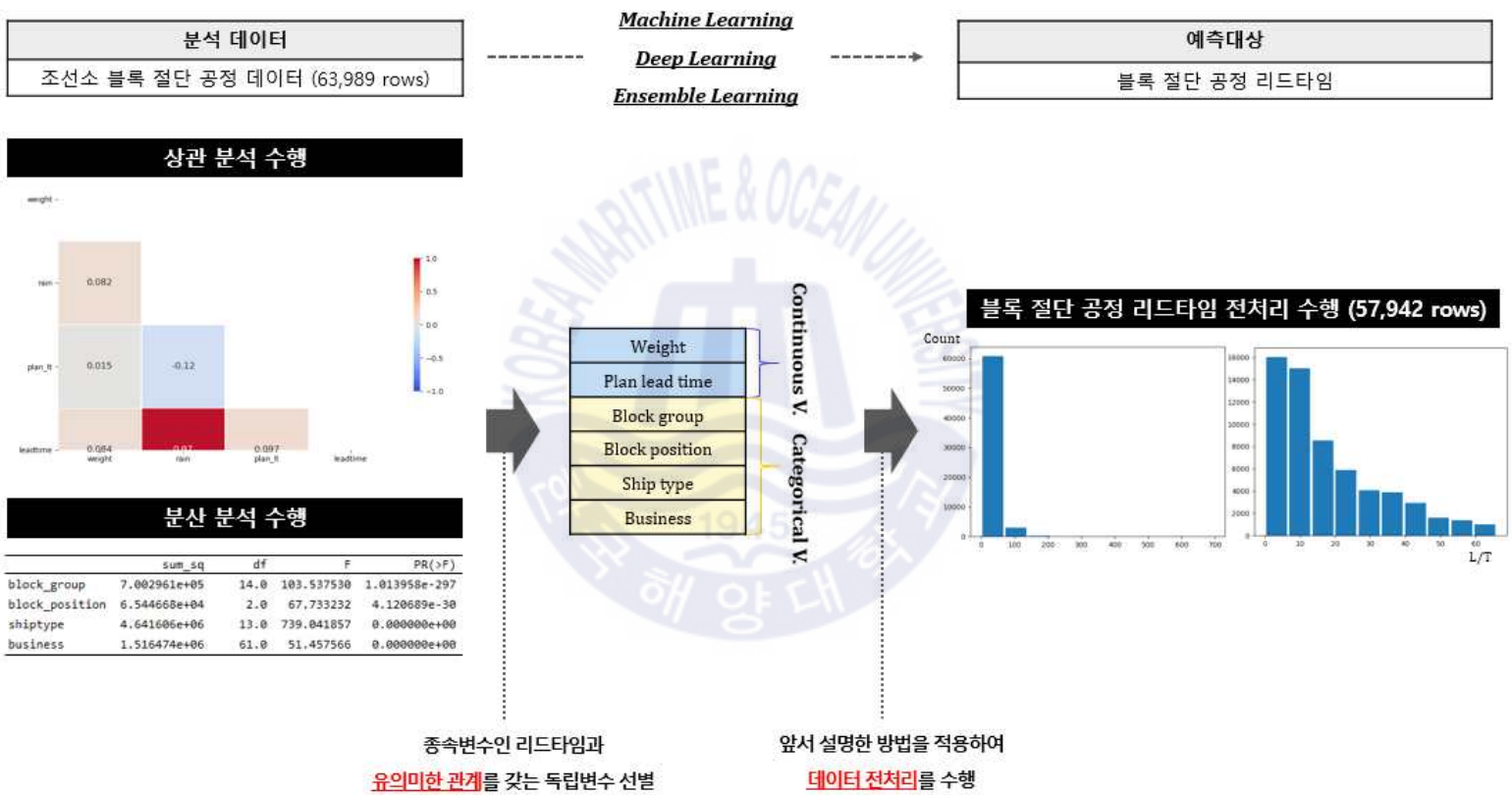


Fig. 24 Block cutting process data analysis process

제 4 장 분석 결과

본 연구에서는 조선 생산 리드타임을 예측하기 위해 수집된 3가지 공정의 데이터를 분석하고자 하며, 각각의 알고리즘에 따른 공정 별 학습 모델의 성능을 여러 가지 평가지표를 통해 확인하고자 한다.

4.1 해양플랜트 배관재 공급망 데이터 분석 결과

가장 먼저 해양플랜트 배관재 공급망의 생산 리드타임 예측 결과를 분석해 보고자 한다. 해양플랜트 배관재 공급망은 제작부터 설치에 이르는 공정으로 이루어지지만 본 연구에서는 제작 공정, 도장 공정의 리드타임을 예측해 보았다. 수집된 데이터에 상관분석, 분산분석을 수행하여 분석에 활용될 독립변수를 선별해 주었으며, 다양한 전처리 기법을 적용하여 데이터 자체에 대한 신뢰도를 높이기 위한 작업을 수행하였다. 본 연구에서는 해양플랜트 배관재 공급망 데이터에 기계학습, 심층학습, 앙상블학습에 해당하는 각각의 알고리즘을 적용하여 예측 모델을 구축하였으며, 다양한 평가지표를 통해 학습 모델의 성능을 확인하였다.

Table 7 Comparing the number of pre-processing data (a)

	Making process	Painting process
Before data preprocessing	32,039 rows	32,039 rows
After data preprocessing	25,590 rows	29,164 rows

데이터 전처리를 수행한 결과를 정리하면 Table 7과 같다. 제작 공정의 경우 32,039개에서 25,590개로 축소된 것을 확인할 수 있으며 도장 공정의 경우 32,039개에서 29,164개로 축소된 것을 확인할 수 있다. 이를 통해 각각의 케이스에 따라 예측 모델 구축을 위한 데이터의 개수가 다르다는 것을 확인할 수 있다.

첫 번째 분석 사례인 해양플랜트 배관재 제작 공정 리드타임의 예측 결과는 Table 8, Fig. 25와 같다.

Table 8 Spool making process data analysis results

	MAE	MAPE	RMSE	RMSLE
regression	7.25	30.91%	9.11	33.21%
decision tree	7.11	26.70%	10.32	36.73%
deep learning	6.42	24.55%	8.53	29.51%
random forest	6.47	23.78%	9.44	32.63%
extra tree	7.23	28.12%	10.52	37.69%
ada boost	9.98	37.07%	13.61	46.95%
gradient boost	7.22	30.38%	9.06	32.77%
xg boost	6.32	25.59%	8.59	30.60%
stacking (D+R+X)	6.29	24.03%	9.00	30.93%
stacking (R+G+X)	6.26	24.32%	8.92	30.72%

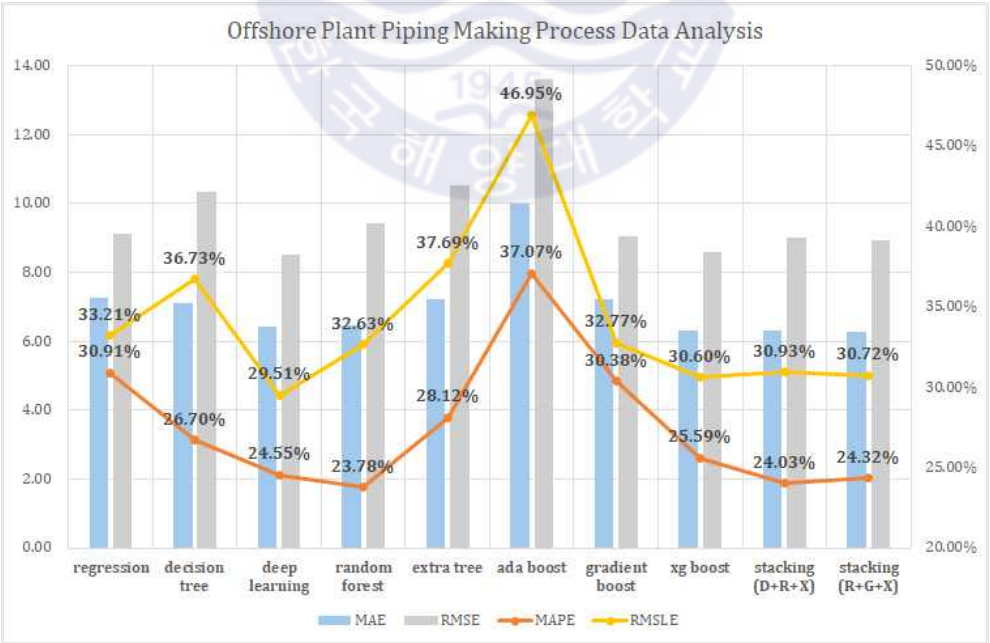


Fig. 25 Spool making process data analysis results

해양플랜트 배관재 제작 공정 데이터에 기계학습, 심층학습, 앙상블학습에 해당하는 알고리즘을 적용한 분석 결과를 비교해 보았을 때, 앙상블학습의 랜덤포레스트 알고리즘을 적용한 모델의 성능이 가장 우수함을 확인하였다.

본 연구에서는 해양플랜트 배관재 제작 공정의 리드타임을 예측하는데 있어서 앙상블학습의 랜덤포레스트 알고리즘을 적용하기 위해 Python의 Scikit-learn 라이브러리를 활용하였다. Scikit-learn은 다양한 기계학습을 수행하기 위한 표준 라이브러리로 데이터 셋 예제, 데이터 전처리 기능 및 다양한 알고리즘을 제공한다는 장점이 있다. 본 연구에서는 해당 데이터의 분석을 수행하기 위해서 Sckit-learn 라이브러리에서 제공하는 ensemble의 RandomForestRegressor 함수를 활용하였으며, 모델을 학습하기 위해 가장 기본적으로 정의하는 파라미터는 n_estimators와 max_features가 있다. n_estimators는 생성할 랜덤포레스트 트리의 개수를 의미하며 결정 트리의 개수가 많을수록 더 깔끔한 dscision boundary가 나오지만, 그 만큼 메모리와 훈련 시간이 증가할 수 있다는 단점이 있다. max_features는 무작위로 선택할 feature의 개수로 max_features 값이 크면 랜덤포레스트의 트리들이 같은 특성을 고려하므로 트리들이 매우 비슷해지고 가장 두드러진 특성을 이용해 데이터에 맞게 예측한다. 반대로 max_features 값이 작으면 랜덤포레스트의 트리들이 서로 매우 달라지므로 과적합이 줄어들 수 있으며, 각 트리는 데이터에 맞추기 위해 깊이가 깊어진다. 그렇기 때문에 적절한 n_estimators와 max_features 값을 설정해 주어야 한다.

같은 알고리즘을 사용하더라도 파라미터에 따라서 성능이 천차만별이므로 연구를 수행함에 있어 데이터에 맞는 적합한 파라미터를 찾는 과정이 매우 중요하다. 그렇기 때문에 가장 좋은 예측 성능을 내는 값을 찾기 위해 수차례 테스트를 해보았으며, 따라서 n_estimators는 500, max_features는 default로 결정하여 최종적으로 분석을 수행하였다. 그 결과 앙상블학습의 랜덤포레스트 알고리즘을 적용한 모델의 성능이 MAPE 23.78%로 가장 우수함을 확인하였다.

두 번째 분석 사례인 해양플랜트 배관재 도장 공정 리드타임의 예측 결과는 Table 9, Fig. 26과 같다.

Table 9 Spool painting process data analysis results

	MAE	MAPE	RMSE	RMSLE
regression	4.47	30.36%	5.63	31.55%
decision tree	4.31	26.32%	6.34	34.48%
deep learning	3.91	25.61%	5.16	28.49%
random forest	3.80	23.34%	5.82	31.37%
extra tree	4.47	27.56%	6.55	35.87%
ada boost	5.19	31.07%	7.14	43.47%
gradient boost	4.47	30.11%	5.63	31.44%
xg boost	4.00	25.53%	5.42	30.10%
stacking (D+R+X)	3.73	23.39%	5.36	29.01%
stacking (R+G+X)	3.72	23.39%	5.39	29.17%

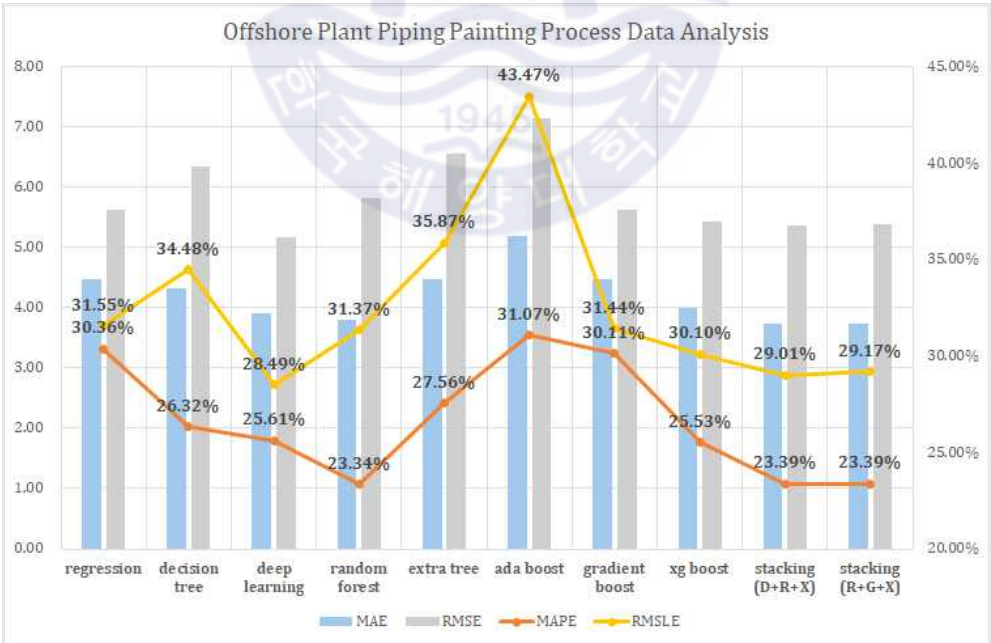


Fig. 26 Spool painting process data analysis results

해양플랜트 배관재 도장 공정 데이터에 기계학습, 심층학습, 앙상블학습에 해당하는 알고리즘을 적용한 분석 결과를 비교해 보았을 때, 제작 공정 데이터의 분석결과와 마찬가지로 앙상블학습의 랜덤포레스트 알고리즘을 적용한 모델의 성능이 가장 우수함을 확인하였다.

본 연구에서는 해양플랜트 배관재 도장 공정의 리드타임을 예측하는데 있어서 앙상블학습의 랜덤포레스트 알고리즘을 적용하기 위해 Python의 Scikit-learn 라이브러리를 활용하였다. 본 연구에서는 해당 데이터의 분석을 수행하기 위해서 Sckit-learn 라이브러리에서 제공하는 ensemble의 RandomForestRegressor 함수를 활용하였다. 또한 모델을 학습하기 위해 가장 기본적으로 정의하는 파라미터인 n_estimators와 max_features 값을 적절하게 설정해주어야 하며, 가장 좋은 예측 성능을 내는 값을 찾기 위해 수차례 테스트해 보았다. 따라서 n_estimators는 500, max_features는 default로 결정하여 최종적으로 분석을 수행하였으며, 그 결과 앙상블학습의 랜덤포레스트 알고리즘을 적용한 모델의 성능이 MAPE 23.34%로 가장 우수함을 확인하였다.

4.2 조선소 블록 조립 공정 데이터 분석 결과

다음으로 조선소 블록 조립 공정의 생산 리드타임 예측 결과를 분석해 보고자 한다. 데이터를 수집한 조선소의 경우, 제품 정보, 공정 정보 그리고 강수량, 강설량과 같은 날씨 정보를 함께 관리하고 있었으며 상대적으로 함께 관리되고 있는 변수의 개수가 많다는 특징이 있었다. 본 연구에서는 조선소 블록 조립 공정 데이터에 기계학습, 심층학습, 앙상블학습에 해당하는 각각의 알고리즘을 적용하여 예측 모델을 구축하였으며, 다양한 평가지표를 통해 학습 모델의 성능을 확인하였다.

Table 10 Comparing the number of pre-processing data (b)

	Block assembly process
Before data preprocessing	11,243 rows
After data preprocessing	10,383 rows

데이터 전처리를 수행한 결과를 정리하면 Table 10과 같으며, 조선소 블록 조립 공정 데이터는 11,243개에서 10,383개로 축소된 것을 확인할 수 있다.

세 번째 분석 사례인 조선소 블록 조립 공정 리드타임의 예측 결과는 Table 11, Fig. 27과 같다.

Table 11 Block assembly process data analysis results

	MAE	MAPE	RMSE	RMSLE
regression	1.50	18.20%	2.02	20.48%
decision tree	1.55	18.36%	2.09	20.56%
deep learning	1.48	17.50%	2.01	19.85%
random forest	1.42	17.36%	1.89	18.99%
extra tree	1.60	19.40%	2.17	21.46%
ada boost	1.40	17.35%	1.82	18.58%
gradient boost	1.58	19.10%	2.08	20.79%
xg boost	1.30	15.62%	1.79	18.12%
stacking (D+R+X)	1.29	15.15%	1.87	18.39%
stacking (D+G+X)	1.29	15.26%	1.86	18.35%

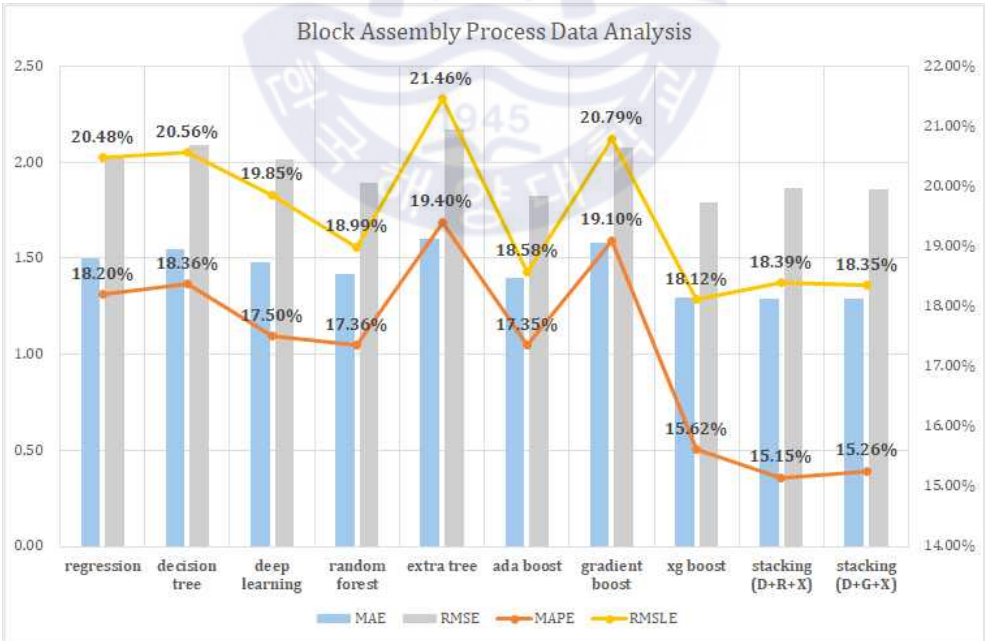


Fig. 27 Block assembly process data analysis results

조선소 블록 조립 공정 데이터에 기계학습, 심층학습, 앙상블학습에 해당하는 알고리즘을 적용한 분석 결과를 비교해 보았을 때, 앙상블학습의 스태킹, 그 중에서도 의사결정나무, 랜덤포레스트, xg boost 알고리즘을 결합한 모델의 성능이 가장 우수함을 확인하였다.

본 연구에서는 조선소 블록 조립 공정의 리드타임을 예측하는데 있어서 앙상블학습의 스태킹 알고리즘을 적용하기 위해 Python의 vecstack 라이브러리를 활용하였다. vecstack은 스태킹을 수행하기 위한 라이브러리로 서로 다른 모델들을 조합해서 새로운 모델을 생성한다. 다양한 알고리즘 조합 중에서도 의사결정나무, 랜덤포레스트, xg boost를 결합한 모델의 성능이 가장 우수하였기 때문에 분석을 수행하기 위해서 Sckit-learn 라이브러리에서 제공하는 tree의 DecisionTreeClassifier 함수, ensemble의 RandomForestClassifier 함수 그리고 xgboost 라이브러리에서 제공하는 XGBClassifier 함수를 함께 사용하였다. 의사결정나무 모델을 학습하기 위해 가장 기본적으로 정의하는 파라미터는 max_depth이다. max_depth는 결정 트리의 최대 깊이로 이를 활용하여 사전 가지치기를 수행하고 과적합을 방지할 수 있으며, max_depth는 20으로 결정하여 분석을 수행하였다. 랜덤포레스트 모델을 학습하기 위해 가장 기본적으로 정의하는 파라미터는 n_estimators와 max_features이며, n_estimators는 100, max_features는 default로 결정하여 분석을 수행하였다. xg boost 모델을 학습하기 위해서 가장 기본적으로 정의하는 파라미터는 max_depth, n_estimators가 있으며, max_depth는 10, n_estimators는 100으로 결정하여 분석을 수행하였다. 이를 결합한 메타모델은 xg boost 알고리즘으로 학습하였으며, 트리를 생성할 때 훈련데이터에서 변수를 샘플링해주는 비율인 colsample_bytree라는 파라미터를 추가해 주었다. 같은 알고리즘을 사용하더라도 파라미터에 따라서 성능이 천차만별이므로 가장 좋은 예측 성능을 내는 값을 찾기 위해 수차례 테스트를 해본 뒤 각 파라미터의 값을 결정하였다. 최종적으로 max_depth는 20, n_estimators는 100, colsample_bytree는 0.5로 결정하여 학습 모델을 구축하였다. 그 결과 앙상블학습의 스태킹, 그 중에서도 의사결정나무, 랜덤포레스트, xg boost 알고리즘을 결합한 모델의 성능이 MAPE 15.15%로 가장 우수함을 확인하였다.

4.3 조선소 블록 절단 공정 데이터 분석 결과

마지막으로 조선소 블록 절단 공정의 생산 리드타임 예측 결과를 분석해 보고자 한다. 데이터를 수집한 조선소의 블록 절단 공정 데이터의 경우, 상대적으로 함께 관리되고 있는 변수의 개수가 적다는 특징이 있었다. 본 연구에서는 조선소 블록 절단 공정 데이터에 기계학습, 심층학습, 앙상블학습에 해당하는 각각의 알고리즘을 적용하여 예측 모델을 구축하였으며, 다양한 평가지표를 통해 학습 모델의 성능을 확인하였다.

Table 12 Comparing the number of pre-processing data (c)

	Block cutting process
Before data preprocessing	63,989 rows
After data preprocessing	57,942 rows

데이터 전처리를 수행한 결과를 정리하면 Table 12와 같으며, 조선소 블록 절단 공정 데이터는 63,989개에서 57,942개로 축소된 것을 확인할 수 있다.

네 번째 분석 사례인 조선소 블록 절단 공정 리드타임의 예측 결과는 Table 13, Fig. 28과 같다.

Table 13 Block cutting process data analysis results

	MAE	MAPE	RMSE	RMSLE
regression	8.80	135.42%	11.88	78.00%
decision tree	6.51	86.03%	10.50	66.94%
deep learning	6.89	103.62%	9.98	65.79%
random forest	6.29	86.60%	9.50	61.55%
extra tree	6.66	87.26%	10.76	67.50%
ada boost	6.65	100.19%	9.97	65.45%
gradient boost	6.49	86.03%	10.45	66.57%
xg boost	6.83	92.08%	10.16	67.48%
stacking (D+G+X)	6.89	72.14%	11.05	69.15%
stacking (D+R+G)	7.55	68.84%	11.98	74.46%

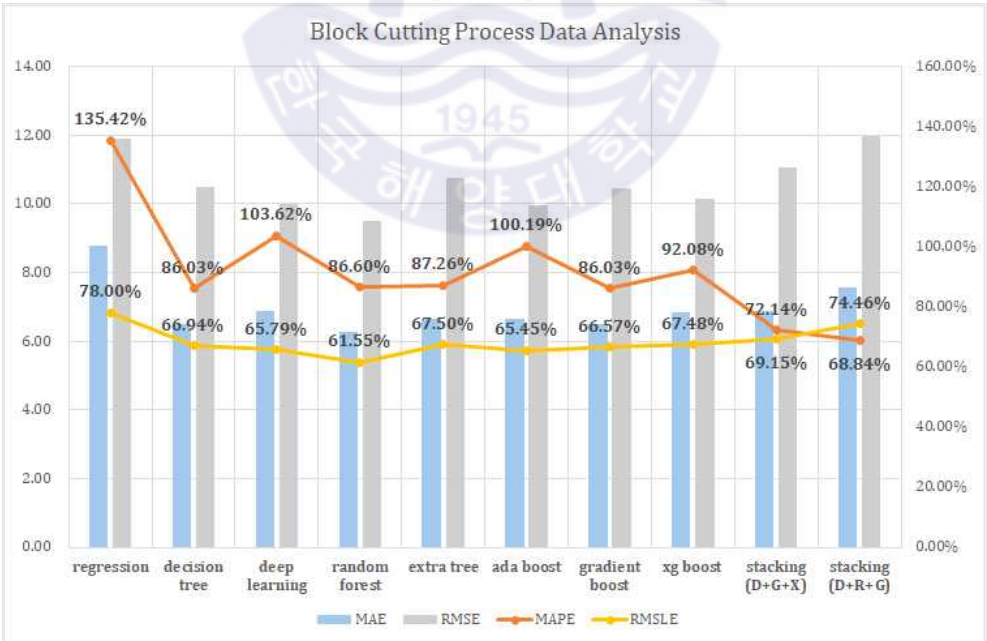


Fig. 28 Block cutting process data analysis results

조선소 블록 절단 공정 데이터에 기계학습, 심층학습, 앙상블학습에 해당하는 알고리즘을 적용한 분석 결과를 비교해 보았을 때, 앙상블학습의 스태킹, 그 중에서도 의사결정나무, 랜덤포레스트, 그래디언트 부스팅 알고리즘을 결합한 모델의 성능이 가장 우수함을 확인하였다.

본 연구에서는 조선소 블록 절단 공정의 리드타임을 예측하는데 있어서 앙상블학습의 스태킹 알고리즘을 적용하기 위해 Python의 vecstack 라이브러리를 활용하였다. 다양한 알고리즘 조합 중에서도 의사결정나무, 랜덤포레스트, 그래디언트 부스팅을 결합한 모델의 성능이 가장 우수하였기 때문에 분석을 수행하기 위해서 Sckit-learn 라이브러리에서 제공하는 tree의 DecisionTreeClassifier 함수, ensemble의 RandomForestClassifier 함수 그리고 ensemble의 GradientBoostingClsaaifier 함수를 함께 사용하였다. 의사결정나무 모델을 학습하기 위해 가장 기본적으로 정의하는 파라미터는 max_depth이며, max_depth는 50으로 결정하여 분석을 수행하였다. 랜덤포레스트 모델을 학습하기 위해 가장 기본적으로 정의하는 파라미터는 n_estimators와 max_features이며, n_estimators는 500, max_features는 default로 결정하여 분석을 수행하였다. 그래디언트 부스팅 모델을 학습하기 위해서 가장 기본적으로 정의하는 파라미터는 max_depth, n_estimators가 있으며, max_depth는 50, n_estimators는 500으로 결정하여 분석을 수행하였다. 이를 결합한 메타모델은 xg boost 알고리즘으로 학습하였으며, max_depth는 50, n_estimators는 500, colsample_bytree는 0.5로 결정하여 최종적으로 학습 모델을 구축하였다. 같은 알고리즘을 사용하더라도 파라미터에 따라서 성능이 천차만별이므로 가장 좋은 예측 성능을 내는 값을 찾기 위해 수차례 테스트를 해본 뒤 각 파라미터의 값을 결정하였다. 그 결과 앙상블학습의 스태킹, 그 중에서도 의사결정나무, 랜덤포레스트, 그래디언트 부스팅 알고리즘을 결합한 모델의 성능이 MAPE 68.84%로 가장 우수함을 확인하였다. 다른 데이터 분석 결과에 비해서 전반적으로 오차율이 높게 나타나는 이유는 실제 데이터에서 함께 고려되고 있는 변수가 적었기 때문에 모델 학습 단계에서 다양한 속성 정보가 반영되지 못했기 때문이라고 판단된다.

본 연구에서는 조선 생산 리드타임을 예측하기 위해 조선소의 공정 데이터 그 중에서도 해양플랜트 배관재 공급망 데이터, 조선소 블록 조립 공정 데이터 그리고 조선소 블록 절단 공정 데이터를 기계학습, 심층학습, 앙상블학습을 적용하여 분석하였다. 분석에는 회귀분석, 의사결정나무, 다층 퍼셉트론, 랜덤 포레스트, 엑스트라 트리, 아다 부스트, 그래디언트 부스팅, xg boost 그리고 각 알고리즘을 결합한 모델인 스태킹까지 약 10가지의 알고리즘을 사용하였으며, 수집된 3가지 공정의 데이터를 분석해 본 결과 모두 앙상블학습에 해당하는 알고리즘을 적용한 모델의 성능이 가장 우수하다는 결론을 얻을 수 있었다. 수집된 데이터의 각 공정에 따라 리드타임 평균 및 편차에 차이가 존재하기 때문에 MAPE 값에 차이가 나타난 것으로 보인다.

제 5 장 리드타임 예측모델 활용 방안

5.1 Python / Simpy를 이용한 이산 사건 시뮬레이션

본 연구에서는 실제 생산 계획 단계에서 예측된 생산 리드타임을 그대로 사용하기 보다는 Fig. 29와 같이 이를 반영한 이산 사건 시뮬레이션을 통해 예측된 리드타임의 적합성을 검증하는 프로세스를 추가할 수 있다. 여기서 이산 사건 시뮬레이션(discrete event simulation, DES)은 이산 사건 시스템을 기반으로 하는 시뮬레이션이다. 이산 사건 시스템은 기준이 되는 시간의 한 시점에서 시스템의 상태를 변화시키는 이벤트가 발생할 때 시스템의 상태가 바뀌는 원리로 구동된다. 이산 사건 시뮬레이션은 정의된 이벤트가 발생하지 않으면 해당 시뮬레이션 모델은 계산을 수행하지 않고 대기 상태를 유지하게 되므로 시스템에서 불필요한 부하가 발생하지 않는다는 장점이 있다.

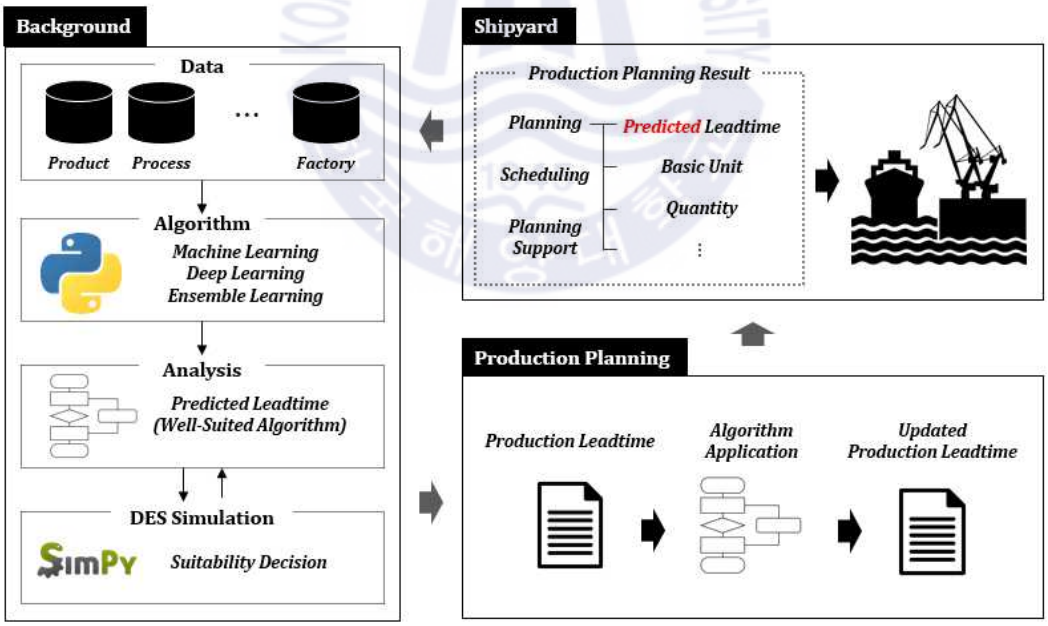


Fig. 29 Suitability decision of prediction lead time with simulation

본 연구에서는 조선소 공정 및 공급망 시뮬레이션을 위해 경량화 DES 엔진을 개발하였다. 이를 위해 Python기반의 Simpy 라이브러리를 활용하였으며, Simpy는 이산 사건 시뮬레이션을 위한 Python 라이브러리이다. 기존의 Simpy 패키지는 독립되어 있는 개별 공정만 구현할 수 있었기 때문에 제조 공정의 선후행 관계 및 병렬 공정 모델링을 위해 Simpy 커널을 customizing 하였다. 이를 그림으로 표현하면 Fig. 30과 같다.

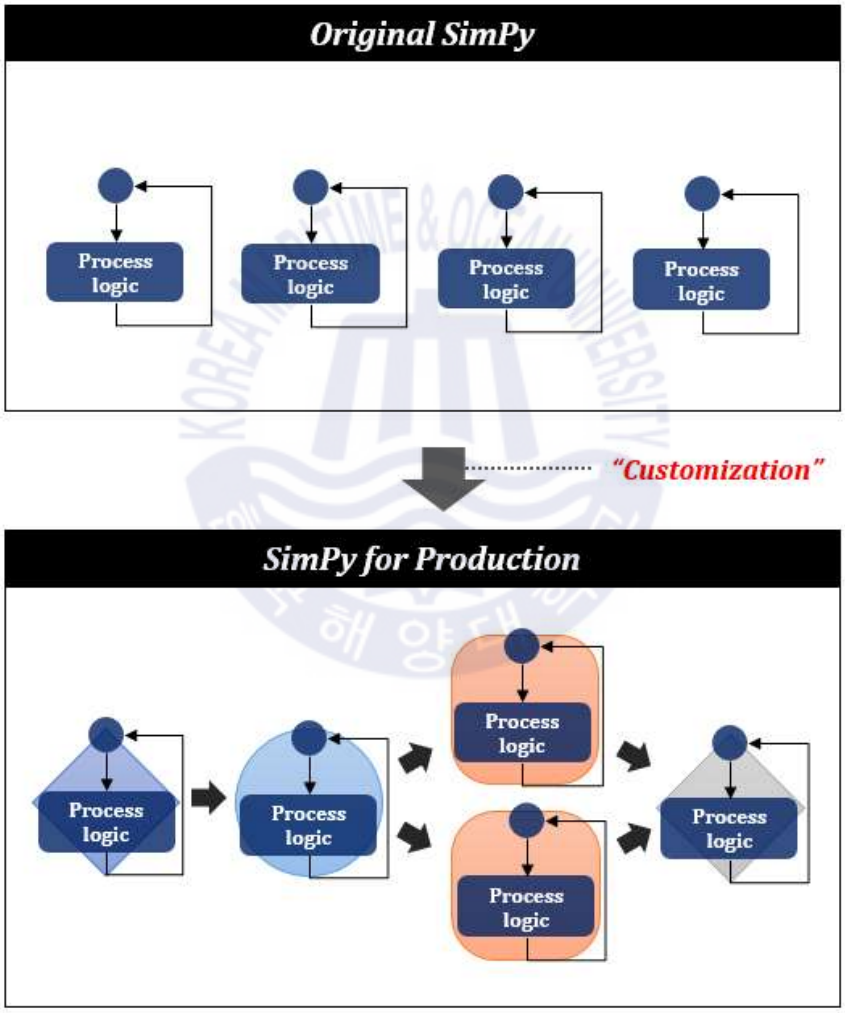


Fig. 30 Simpy kernel customizing

5.2 시뮬레이션 대상 선정

시뮬레이션은 해양플랜트 배관재 제작 및 도장 공정으로 Fig. 31과 같이 각 공정의 협력사를 대상으로 수행하였다. 시뮬레이션 모델을 구축하기 위해 사용한 데이터는 spool number, 제품 정보, 공정 정보, 협력사 정보, 계획 리드타임, 예측 리드타임, 실제 리드타임으로 구성되어 있다. 해당 데이터를 기반으로 시뮬레이션을 수행할 때, Simpy 개발을 통해서 기존에 약 10분 이상 소요되던 시뮬레이션을 5초 이내로 줄여 볼 수 있었다.

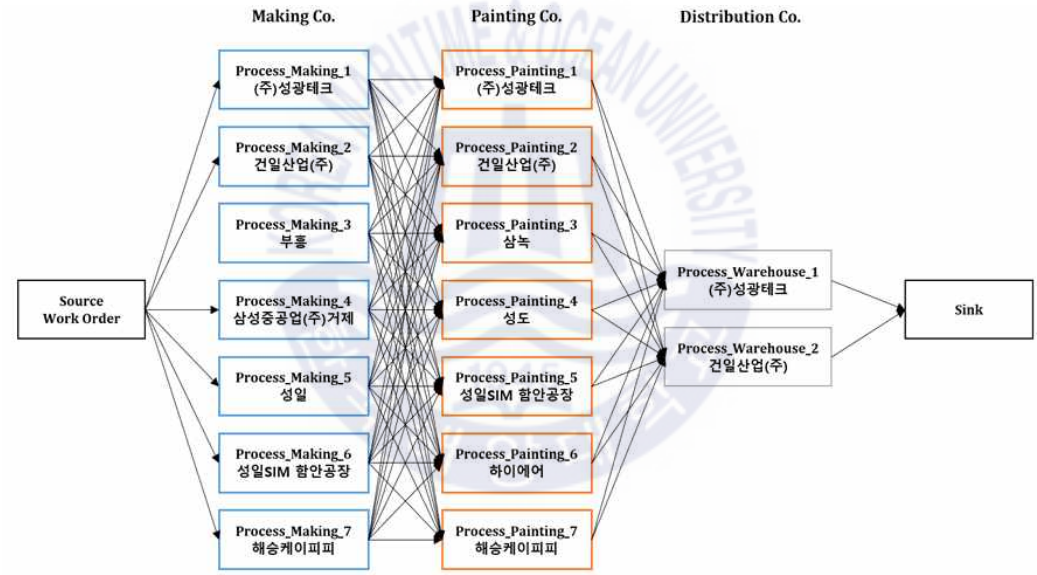


Fig. 31 Selection of simulation targets

5.3 시뮬레이션 적용 및 분석

계획 리드타임, 예측 리드타임, 실제 리드타임을 반영한 해양플랜트 배관재 제작 및 도장 공정의 시뮬레이션 결과를 비교해 볼 수 있다. 전체 공정의 리드타임, 협력사 별 전체 대기시간, 각 공정의 협력사 별 작업시간 그리고 각 공정의 협력사 별 평균 spool의 개수를 비교해 보았으며, 결과는 Fig. 32 ~ 37과 같다.

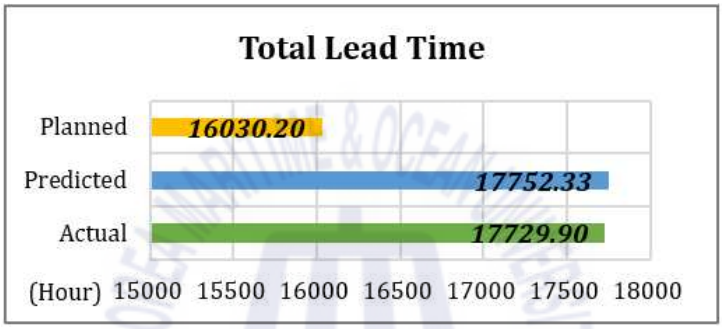


Fig. 32 Comparison of the entire process lead time

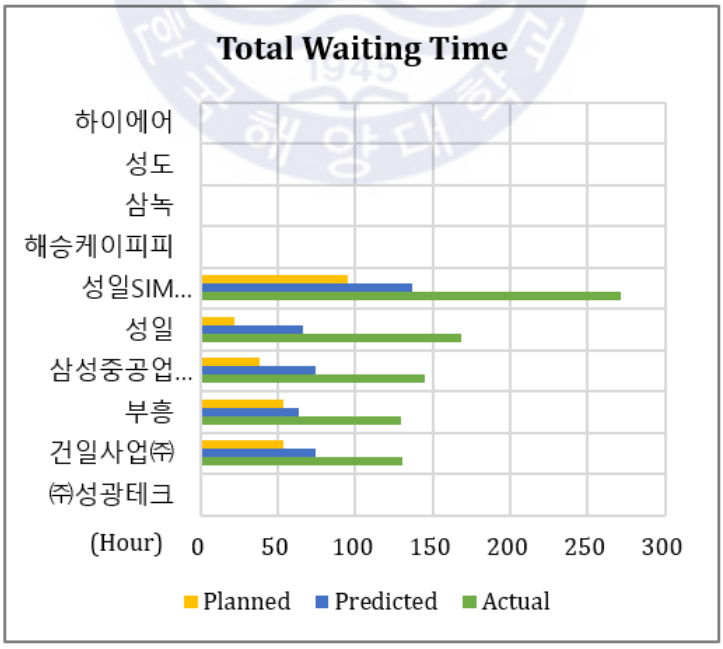


Fig. 33 Comparison of overall waiting time

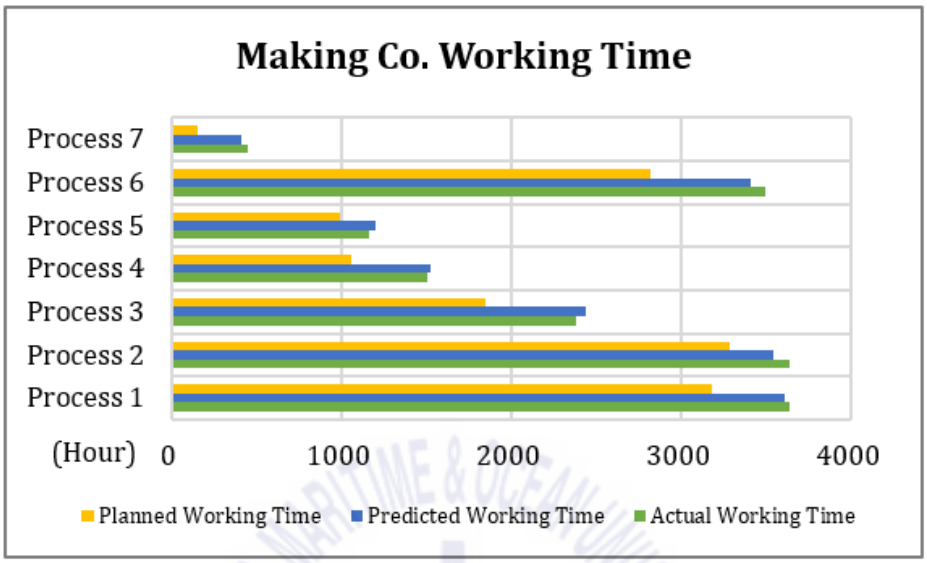


Fig. 34 Comparison of working time in spool making process

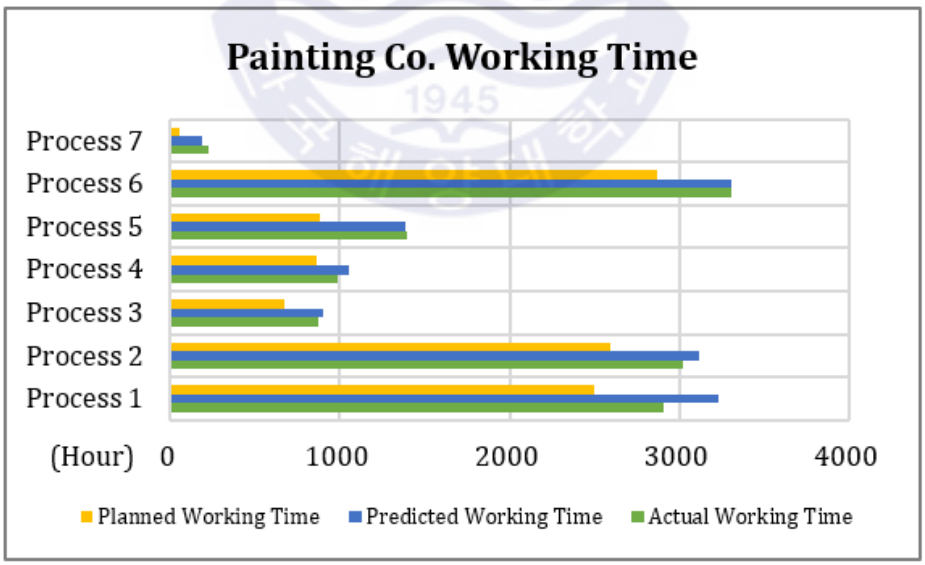


Fig. 35 Comparison of working time in spool painting process

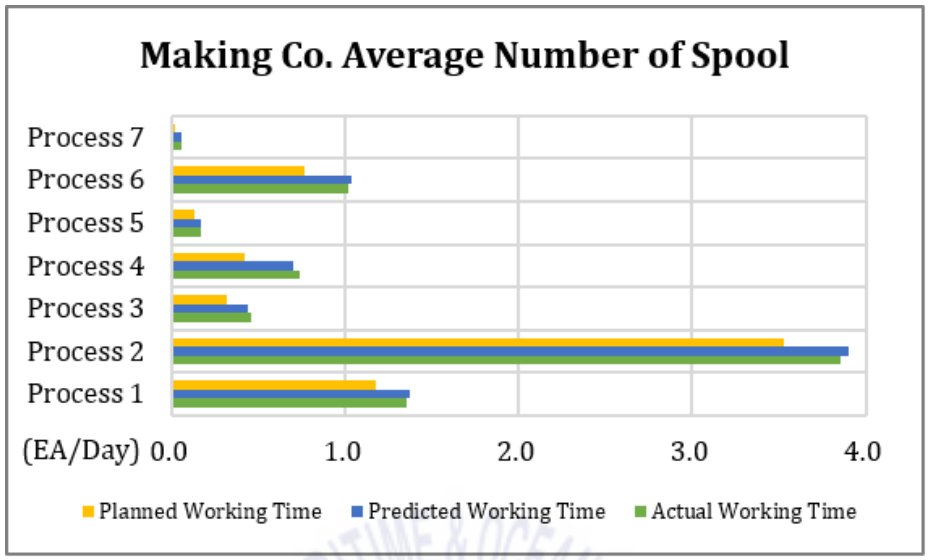


Fig. 36 Comparison of average number of spools by making co.

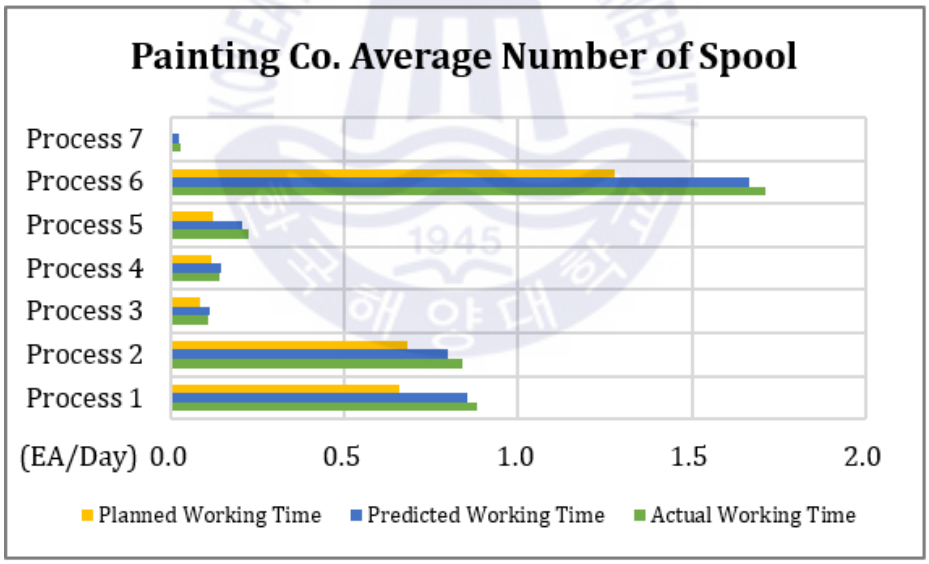


Fig. 37 Comparison of average number of spools by painting co.

결과적으로 전체 공정의 리드타임, 협력사별 전체 대기시간, 각 공정의 협력사별 작업시간 그리고 각 공정의 협력사별 평균 spool의 개수를 비교해 보았을 때, 계획 정보보다 예측 정보가 실제 정보와 유사함을 확인할 수 있었다.

제 6 장 결론

6.1 연구 결론

본 연구에서는 조선소의 생산 리드타임 기준정보 관리를 위해 빅데이터 분석 방법론을 적용하였다. 조선소에서 관리되는 의장, 조립, 강재 절단 공정에 해당하는 해양플랜트 배관재 공급망 데이터, 조선소 블록 조립 공정 데이터, 조선소 블록 절단 공정 데이터를 수집하였으며, 각 공정의 생산 리드타임을 개선하기 위한 방법론을 적용하였다. 제품 정보, 공정 정보 등 다양한 속성 정보를 반영한 기계학습, 심층학습, 앙상블학습 예측 모델을 생성하였다. 수집된 여러 공정의 데이터를 분석해 본 결과, 앙상블학습을 적용한 예측 모델의 성능이 가장 우수하다는 결론을 얻을 수 있었다. 본 연구를 통해서 생산 리드타임을 예측하는데 있어서 빅데이터 분석, 그 중에서도 앙상블학습의 적용가능성을 제시해 보고자하며 표준 리드타임 대비 예측 리드타임을 통한 생산 리드타임 기준정보의 체계적인 관리를 가능하게 하고자 한다. 또한 본 연구의 결과로 *IJNAOE (International Journal of Naval Architecture and Ocean Engineering)*에 Machine Learning Methodology for Management of Shipbuilding Master Data 라는 제목으로 논문 게재를 신청하여 현재 리뷰 중에 있다. 향후, 다양한 학습 알고리즘이 기업의 백그라운드에서 돌아가며 연속적으로 실적과 비교되어 가장 적합한 예측 알고리즘을 추천하는 시스템을 설계하고자 한다.

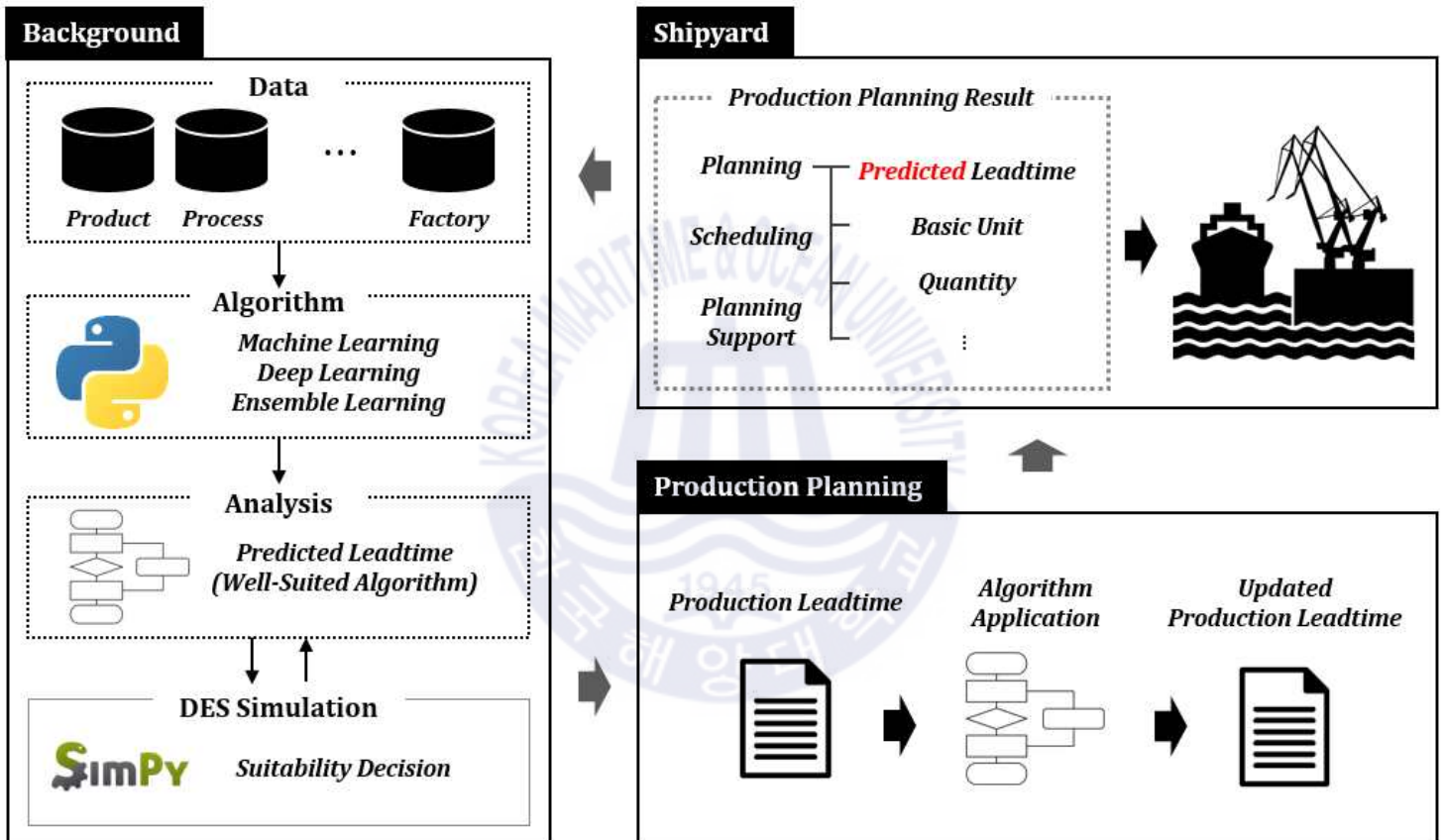


Fig. 38 System configuration for the application of prediction model

Reference

김기열, 2017. 치의학 연구에서 이상치의 처리. *대한치과의사협회지*, 55(9), pp.604-616.

김민석, 백준걸, 2012. 앙상블 학습을 이용한 DRAM모듈 출하 품질보증 검사 불량 예측. *산업공학(IE interfaces)*, 25(2), pp.178-186.

김성훈, 노명일, 김기수, 2016. 조선 해양 산업에서의 응용을 위한 하둡 기반의 빅데이터 플랫폼 연구. *한국CDE학회 논문집*, 21(3), pp.334-340.

김영주 등, 2013. 선박설계 자동화를 위한 빅데이터 기술 및 분석기법 연구. *한국통신학회 학술대회논문집*, pp.213-215.

김지혜, 2018. 조선 생산 리드타임 예측을 위한 기계학습 방법론에 관한 연구. 석사학위논문. 부산:한국해양대학교.

민성환, 2014. 개선된 배깅 앙상블을 활용한 기업부도예측. *Journal of Intelligence and Information Systems*, 20(4), pp.121-139.

박선, 정민아, 이성로, 2012. 앙상블 학습을 이용한 적조 발생 예측의 성능향상. *전자공학회논문지-SP*, 49(1), pp.41-48.

양선모, 이순요, 이석주, 1992. 조선업의 CIM 시스템을 위한 생산정보 시스템의 구축방안에 관한 연구. *대한전자공학회 학술대회*, pp.806-810.

이동하, 박재훈, 배혜림, 2013. 조선 산업에서 프로세스 마이닝을 이용한 블록 조립 프로세스의 계획 및 실적 비교 분석. *한국전자거래학회지*, 18(4), pp.145-167.

이병우, 양지훈, 2013. 지역 전문가의 앙상블 학습. *한국정보과학회 학술발표논문집*, 35(1A), pp.120-121.

이상헌, 이상형, 양지현, 2014. 기계 학습의 앙상블 기법을 이용한 운전자 차선 변경 의도 예측. *한국CDE학회 학술발표회 논문집*, pp.493-495.

함동균, 2016. 조선소 의장품 조달관리를 위한 데이터마이닝 방법론에 관한 연구. 석사학위논문. 부산:한국해양대학교.

Bibliography

김경민 등, 2013. 불균형 데이터 처리를 위한 과표본화 기반 앙상블 학습 기법. *한국정보과학회 학술발표논문집*, pp.686-688.

신한솔, 박철수, 2016. 랜덤 포레스트 모델을 위한 데이터 전처리 기법의 적용. *대한건축학회 학술발표대회 논문집*, pp.633-634.

이재구, 이태훈, 윤성로, 2014. Big Data 분석을 위한 Machine Learning. *한국통신학회지(정보와통신)*, 31(11), pp.14-26.

장영재, 2012. 제조 분야에서의 빅데이터 기술 활용. *한국통신학회지(정보와통신)*, 29(11), pp.30-35.

조성준, 강석호, 2016. 머신러닝(인공지능)의 산업 응용. *Industrial Engineering Magazine*, 23(2), pp.34-38.

Hur, M.H. et al., 2015. A study on the man-hour prediction system for shipbuilding. *Journal of Intelligent Manufacturing*, 26(6), pp.1267-1279.

Appendix

본 논문의 이해를 돕기 위해 연구에 사용된 데이터의 변수에 대한 설명을 추가하였다.

Table 14 Description of the variables in the spool procurement process

Column명	실제Column명	의미
DIA	DIA	배관재의 직경 정보
Length	길이(mm)	배관재의 길이 정보
Weight	중량	배관재의 중량 정보
Member Count	부재 수	연결된 부재 수
Joint Count	Joint수	연결된부재간의Joint수
Emergency	긴급 여부	배관재 제작의 우선 순위
Apply Lead time	적용L/T	긴급 여부와 관련(정상, 준긴급, 긴급)
STG	STG	배관재설치Stage
Service	Service	배관재에 흐르는 유체의 종류
Pass	관통	관통여부(Above,Bottom,Penetration)
Sch	Sch	Schedulenumber,배관재의두께
Making_Co	제작협력사	배관재 제작 협력사
After2_Co	후2협력사	배관재 도장 협력사

Table 15 Description of the variables in the block assembly process

Column명	실제Column명	의미
Weight	중량(TON)	블록의 중량
Length	Length (M)	블록의 길이
Breadth	Breadth (M)	블록의 폭
Height	Height (M)	블록의 높이
Plan lead time	계획L/T	생산 계획 단계에서 계획 된 리드타임
Rain	강수량	작업 기간 동안의 강수량
Snow	강설량	작업 기간 동안의 강설량
Team	팀	잡업을 수행한 팀 (조립 1팀, 조립 2팀)
Process	세부공정	블록 조립의 세부 공정 (대조립, 중조립)
Ship type	선종	블록이 탑재 될 선종
Project	프로젝트	프로젝트 번호
Block	블록	블록의 이름
Assembly	Ass’y	조립된 블록의 이름
PCG	PCG	Productivity Control Group, 생산성관리단위
PCG name	PCG명	블록이 설치 될 선체 내 구역
Business	업체명	협력사

Table 16 Description of the variables in the block cutting process

Column명	실제Column명	의미
Weight	중량	블록의 중량
Plan leadtime	Workday_Plan	생산 계획 단계에서 계획 된 리드타임
Block group	블록 그룹	블록의 그룹
Block position	블록 위치	블록의 위치
Ship type	선종	블록이 탑재 된 선종
Business	실적업체	협력사