Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2023.46.1.023

Vol. 46, No. 1 : 23-31, March 2023
ISSN : 2005-0461 / EISSN : 2287-7975

# Machine Learning Algorithm for Estimating Ink Usage

Se Wook Kwon · Young Joo Hyun · Hyun Chul Tae†

Department of Digital Healthcare Research Korea Institute of Industrial Technology

# 머신러닝을 통한 잉크 필요량 예측 알고리즘

권세욱 · 현영주 · 태현철†

한국생산기술연구원 디지털헬스케어연구부문

Research and interest in sustainable printing are increasing in the packaging printing industry. Currently, predicting the amount of ink required for each work is based on the experience and intuition of field workers. Suppose the amount of ink produced is more than necessary. In this case, the rest of the ink cannot be reused and is discarded, adversely affecting the company's productivity and environment. Nowadays, machine learning models can be used to figure out this problem. This study compares the ink usage prediction machine learning models. A simple linear regression model, Multiple Regression Analysis, cannot reflect the nonlinear relationship between the variables required for packaging printing, so there is a limit to accurately predicting the amount of ink needed. This study has established various prediction models which are based on CART (Classification and Regression Tree), such as Decision Tree, Random Forest, Gradient Boosting Machine, and XGBoost. The accuracy of the models is determined by the K-fold cross-validation. Error metrics such as root mean squared error, mean absolute error, and R-squared are employed to evaluate estimation models' correctness. Among these models, XGBoost model has the highest prediction accuracy and can reduce 2134 (g) of wasted ink for each work. Thus, this study motivates machine learning's potential to help advance productivity and protect the environment.

Keywords : Machine Learning, XGBoost, Ink Usage, Regression

## 1. 서 론

ESG는 환경(Environmental), 사회(Social), 지배구조(Governance)를 조합한 단어로 최근 전  계적으로 ESG 경영에 관심이 높아지고 있다[7]. 패키징 프린팅 분야에서도 ESG 경영을 위해 지속가능한 인쇄(Sustainable Printing)에 관심이 높아지고 있다[10]. 지속 가능한 인쇄를 위한 해결책은 환경 친화적인 잉크 사용부터 재활용, 작업 흐름에서 폐기물 처리에 이르기까지 다양한 분야를 망라하고 있다. 하지만 프린팅 패키징 산업의 규모가 비교적 작아 환경문제를 다루는데 얼마나 많은 기업들이 참여를 할 것인지는 의문이다. 이익과 환경 사이에 갈등이 있을 때 대부분 환경적 책임을 지는 척하며 이익을 추구하는 방식을 택할 것이다. 본 연구는 패키징 프린팅 기업의 수요에 의해 환경과 이익 두 가지를 모두 해결할 수 있는 방안을 제시한다.

기존 패키징 인쇄 산업에서는 잉크 제조 방식을 현장 작업자의 경험과 직관을 기반으로 하고 있다. 하지만 작업자의 의해 진행된 중요한 의사결정(잉크 배합비율, 스케줄링, 잉크 생산량)이 조금이라도 틀릴 경우 기업의 생산성에 악영향을 미친다. 예를 들어, 잉크를 필요량 보다 과다하게 생산하게 되는 경우 남는 잉크는 재사용 될 수 없어 폐기되어 생산비용이 올라가고 환경에 악영향을 미

Received 21 November 2022; Finally Revised 26 December 2022; Accepted 27 December 2023
† Corresponding Author : sage@kitech.re.kr

친다. 반면 잉크를 부족하게 생산한 경우 잉크를 재생산할 때까지 작업이 중단되기에 생산 비용이 올라가므로 적정 여유 분의 잉크만 남도록 잉크 사용량의 정확한 예측이 필요하다.

필요 잉크량을 정확히 계산해 필요한 만큼 만드는 것은 비용 절감, 생산 효율성 향상, 그리고 환경 보호를 위한 윈-윈 솔루션이 될 수 있다. 보통 프린팅 작업에서 잉크 코스트는 평균 2%~5%의 비용[19]을 차지한다. 전체 작업에서는 적은 부분을 차지하는 것 같지만 대량 인쇄를 할 경우 잉크 코스트는 무시하지 못할 수준이다. 과거 1980년부터 1990년 사이 Procter & Gamble Co.는 생산 비용을 줄이기 위해 적은 잉크를 사용하는 전략을 사용했다. 이 전략은 실제 매년 약 200만 달러의 비용 절감을 하는 효과를 가져왔다[6].

기존 잉크 필요량 예측과 관련된 선행 연구는 패키징 프린팅 산업의 규모가 작아 거의 진행되지 않았다. 패키징 프린팅 산업에서 대략적으로 잉크 필요 예측량을 계산하기 위해 사용되고 있는 SPANKS식[13]은 종이 종류(Stock), 인쇄 방식(Process), 면적(Area), 매수(Number), 이미지 종류(Kind of Image), 잉크 비중(Specific Gravity of Ink)으로 이루어진 간단한 수리식이다. 하지만 SPANKS식과 같은 경우 3장에서 기술할 현실적인 패키징 프린팅 작업의 특성을 다 반영하지 못해 잘못된 예측을 하기 쉬우며 현장에서도 숙련된 경력자의 직관을 기반으로 필요 잉크량을 예측하여 제조하는 상황이다.

패키징 프린팅 산업에도 인공지능 기반 조색 시스템을 도입하면 앞서 설명한 문제가 해결되며 폐기되는 잉크가 줄어 기업의 이익 추구와 환경 보호의 의무 또한 충족할 수 있다. 본 연구는 4차 산업혁명 전환 시점에 맞추어 잉크 생산량 예측 머신러닝 알고리즘 관련 연구를 진행하였다.

머신러닝은 인공지능(artificial intelligence)의 일종으로 컴퓨터에 명시적인 프로그램 없이 배울 수 있는 능력을 제공하는 분야로 정의되며 인간이 학습하듯 컴퓨터에도 정보들을 제공하여 스스로 경험하고 학습하게 함으로써 새로운 지식을 창조하는 것을 말한다[11]. 머신러닝 방법론은 대표적으로 지도학습(Supervised Learning)과 비지도 학습(Unsupervised Learning)으로 구분된다.

지도학습은 각 객체 별로 속성에 대한 입력 벡터와 목표치(Label)가 주어진 학습데이터(Training Data)가 존재할 때, 이로부터 하나의 함수를 유추하여 미래의 새로운 객체에 대한 목표치를 예측하는 방법이다. 지도학습은 주로 예측 모델링(Predictive Modeling)에 활용되며, 목표치의 형태에 따라서 분류(Classification)와 회귀(Regression) 등이 있다. 본 연구에서는 지도학습 중 회귀모델을 통해 잉크 필요량을 예측하고자 한다.

대표적인 회귀모델로는 다중선형회귀(Multiple Linear Regression)[17], 의사결정나무(Decision Tree)[15], 랜덤포레스트(Random Forest)[22], GBM(Gradient Boosting Machine)[8], XGBoost(eXtreme Gradient Boosting)[5] 알고리즘이 있다.

잉크 필요량 예측 알고리즘 모델은 인쇄하는 종이의 면적, 색상 망점 비율(Halftone Dot), 인쇄매수, 잉크 종류, 종이의 앞/뒷면, 종이 종류의 정보를 이용한다. 현실적인 인쇄의 특성을 다 반영하기 위해서 복잡한 데이터를 사용하지만 기계학습(Machine Learning)의 Computing Power를 이용하면 복잡하고 방대한 양의 정보라도 쉽게 처리 가능하도록 만들어 준다.

본 연구의 목표는 대표 머신러닝 모델들을 비교하며 잉크 필요 생산량 예측에 가장 효과적인 모델을 탐색한다.

## 2. 선행 연구

최근 다양한 산업분야에서 머신러닝을 활용하여 보다 효율적인 의사결정을 하기위한 연구가 진행되고 있다.

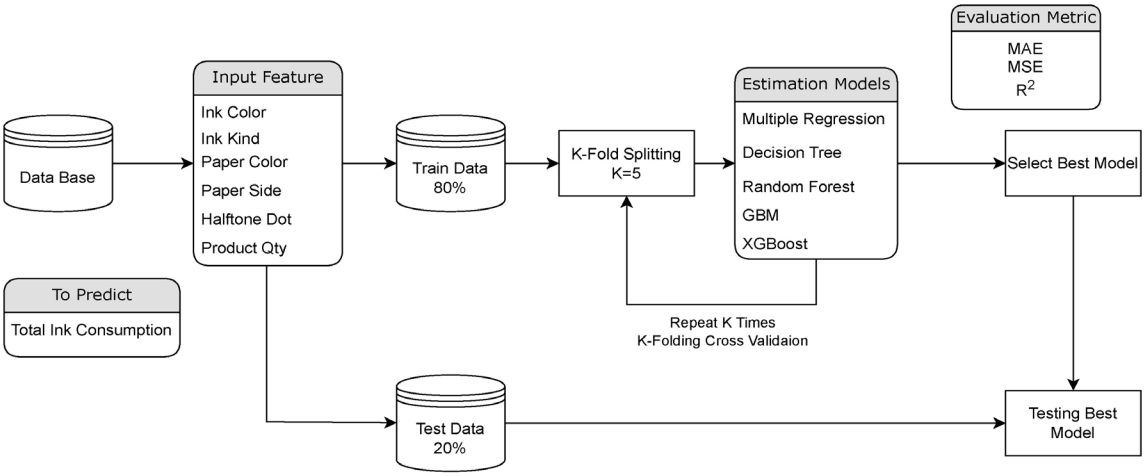
Gkerekos[9]는 선박 엔진에 들어가는 연료 필요량을 머신러닝을 통해 예측하여 선박 관리의 효율성과 이윤을 향상시키기 위한 연구를 진행하였다. 다중선형회귀 모델부터 의사결정나무 모델들까지 다양한 모델들을 R-squared, MAE, MSE 등 다양한 평가지표를 사용하여 비교하여 가장 성능이 좋은 모델을 사용하였다. 또한 가장 성능이 좋은 모델의 Hyper Parameter를 튜닝하며 성능을 한차례 더 개선하였다.

Shehadeh[20]는 중장비의 수명예측을 위해 중장비 수명예측에 있어 중요한 연속형(Numerical)변수와 범주형(Categorical) 변수들을 선별하고 범주형(Categorical) 변수들을 Input으로 활용할 수 있는 CART(Classification And Regression Tree)[14]기반의 Modified Decision Tree, LightGBM, XGBoost 모델들의 MAE, MSE, MAPE, R-squared를 비교하였다.

Jung and Kim[12]은 철도 분야 고장진단 모델에 다양한 고장진단 파라미터들을 사용하여 트리기반의 머신러닝 알고리즘들을 비교하였다. 그 중 트리 기반의 Random Forest 모델이 가장 좋은 성능을 보였다.

Razi[18]는 데이터가 비선형성관계를 가지며 범주형(Categorical) 변수들을 포함하고 있을 때 Neural Networks 모델과 CART 기반의 모델들의 우수성에 대해 연구했다.

과거 연구들과 같이 다양한 산업분야에서 머신러닝을 통해 효율적인 의사결정을 위한 연구가 진행되고 있다. 위 연구들의 산업분야는 모두 다르지만 예측하고자 하는 Target 변수를 가장 잘 설명할 수 있는 머신러닝 모델을 찾아내기 위해 다양한 모델들을 비교하였다. 또한 Target




<Figure 1> Visual Representation of the Suggested Methodology

변수를 결정짓는데 중요한 인자로 작용하게 될 변수들을 도메인 지식(Domain Knowledge)을 활용하여 모델의 학습 변수로 활용하였다.

본 연구에서는 아직 머신러닝 도입 선례가 없는 패키징 프린팅 분야에서 작업에 필요한 잉크 필요량을 계산하기 위한 머신러닝 모델을 연구한다. 선형 회귀 모델인 다중선형회귀(Multiple Linear Regression) 모델과 변수들 간의 비선형성을 가지고 명목형 변수를 사용할 때 우수한 성능을 보이는 CART기반의 머신러닝 알고리즘들을 비교한다.

## 3. 데이터 및 연구 방법

### 3.1 패키징 생산 공정 작업방식

현장의 조색 작업 공정 방식은 다음과 같다.

(1) 고객 주문 시안 접수: 고객이 원하는 패키징 인쇄물의 디자인 시안을 전달받는 단계로 각 색상의 LAB Value와 같은 구체적인 품질 기준을 협의하는 과정이다.

(2) 잉크 조색: 대량 인쇄를 위해 고객의 주문량을 인쇄하기 위해 필요한 잉크를 만드는 과정으로 재고로 보유하고 있는 색이 있는 경우를 제외하고는 새롭게 여러 잉크를 섞음(Mix)으로써 원하는 색의 잉크를 조색하는 과정이다. 또한 고객의 주문량과 고객 시안의 색별 면적 등을 고려해서 어느 정도의 잉크를 새로 만들어야 하는지 의사결정 하는 단계이다.

(3) 생산 시안 디자인: 대량 생산이 가능하도록 고객으로부터 받은 디자인 시안을 생산 시안으로 재 디자인하는 과정으로 Off-Set 인쇄를 위한 PS판의 망점 설계 과정을 포함한다.

(4) 시험 인쇄: 조색 된 잉크에 따라 설비를 통해 시험 생산하는 단계로 기계의 예열을 위해서 적정 매수(50장 이상)를 인쇄하는 단계이다.

(5) 색상 검사: 시험 인쇄된 인쇄물이 고객이 주문한 색상과 허용범위 내로 일치하는지 검사하는 단계로 일치할 경우 대량 인쇄로 넘어가며 그렇지 않은 경우 색상 불일치 요인을 찾기 위해서 이전의 모든 과정(생산 시안 디자인, 잉크 조색)을 재작업을 실시한다.

(6) 대량 인쇄: 고객이 주문한 주문량에 맞게 인쇄물을 대량 생산하는 과정이다.

### 3.2 사용 데이터셋

본 연구에서 사용한 데이터셋은 국내 패키징 인쇄 기업인 덕수산업으로부터 2021-04-01부터 2021-11-03까지 수집된 1,034개의 데이터이다. 데이터의 변수로는 인쇄 매수(Product Qty), 한 면에 해당 잉크의 면적 비율인 망점 비율(Halftone Dot), 작업자가 생산한 잉크 총량(조색양), 작업자가 작업 후 남은 잉크량, 마지막으로 조색양에서 남은 잉크량을 빼서 계산한 작업에 사용한 잉크량(Ink Used)으로 구성되어 있다. 또한 종이와 잉크의 물성치를 반영할 수 있도록 잉크의 색(Ink Color), 잉크의 종류(Ink Kind), 종이의 색(Paper Color), 종이의 종류(Paper Kind), 종이의 앞/뒤면(Paper Side)의 데이터를 포함한다. 데이터 행은 총 1,034개이며 <Table 1>은 머신러닝 Input Feature와 Target Feature를 나타낸다.

작업에 총 필요한 잉크량을 Total Ink Consumption이라 한다면 작업 한 장에 들어가는 잉크량(One Printing Ink Consumption)에 작업에 필요한 매수(Printing Quantity)를 곱해 구할 수 있다.

<Table 1> Data Feature Description

| Data Feature | Data Description |
|---|---|
| Input Feature | |
| Ink Color | Color code of ink (Categorical Feature) |
| Ink Kind | Divided into regular/UV ink<br>UV uses more ink for the same operation (Categorical Feature) |
| Paper Kind | Type of paper used for the job (Categorical Feature) |
| Paper Color | Color of the paper used for the job (Categorical Feature) |
| Paper Side | The front/back of the cover to print (Categorical Feature) |
| Paper Size | Area of paper ($m^2$) |
| Halftone Dot | The percentage of ink in paper area (%) |
| Product Qty | Number of papers required for the job |
| Target Feature | |
| Ink Used | Amount of ink used by the job (g) |

***Total Ink Consumption (g) = Single Printing Ink Consumption (g)× Printing Quantity***

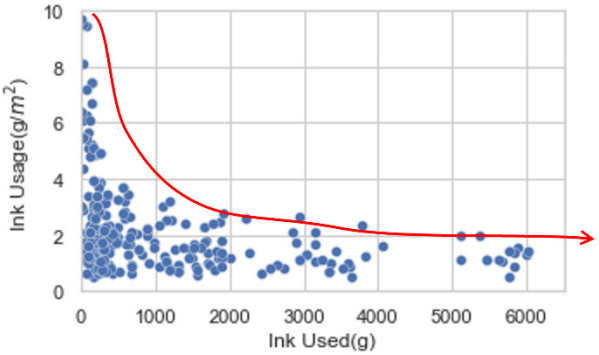
작업 한 장에 들어가는 잉크량(Single Printing Ink Consumption)은 종이 한 장의 크기(Paper Size)에 종이에서 잉크가 차지하는 면적 비율(Halftone Dot)을 곱하면 한 페이지를 차지하는 잉크 면적을 구할 수 있다. 한 페이지에 칠해지는 잉크 면적에 단위 면적 $1m^2$당 필요한 잉크량인 Ink Usage($g/m^2$)를 곱하면 작업 한 장에 들어가는 잉크량(Single Printing Ink Consumption)을 구할 수 있다. Ink Usage($g/m^2$)는 작업에 사용하는 종이와 잉크의 물성치에 영향을 받는다. 본 연구에서는 Ink Usage($g/m^2$)에 영향을 줄 수 있는 변수로 잉크의 색(Ink Color), 잉크의 종류(Ink Kind), 종이의 색(Paper Color), 종이의 종류(Paper Kind), 종이의 앞/뒤면(Paper Side)을 사용한다.

***Single Printing Ink Consumption(g) =Paper Size ($m^2$)× Halftone Dot (%) × Ink Usage ($g/m^2$)***

Ink Usage($g/m^2$)에 영향을 줄 수 있는 변수가 많은 만큼 Ink Usage($g/m^2$) 값을 정량적으로 나타내기에 어려움이 있다. 그러므로 본 연구에서는 머신러닝 모델 중 명목형 변수(Categorical Feature)와 변수간 비선형성을 표현하는데 적합한 CART(Classification And Regression Tree)[14] 기반의 알고리즘들인 의사 결정나무, 랜덤 포레스트, GBM, XGBoost를 사용하여 선형회귀 모델인 Multiple Regression Model[23]과 비교한다.

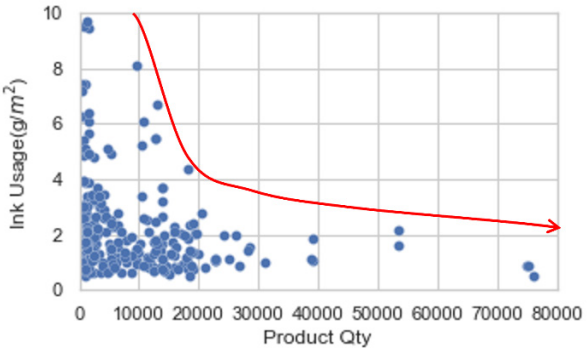
## 3.3 데이터 EDA(Exploratory Data Analysis)

본 데이터셋에서 단위면적 $1m^2$ 당 필요한 잉크량을 Ink Usage($g/m^2$)라고 정의하다. 실제 패키징 생산 공정에서는 3.1. (4) 시험 인쇄 단계에서 테스트 작업을 위해 잉크를 소모하게 된다. 그리고 실제 인쇄 작업에서 인쇄기에 잉크를 넣을 시 잉크 탱크에 저장되고 잉크 펌프를 통해 이동되는 과정에서 펌프와 탱크 벽면에 붙는 최소한의 잉크(Minimum Circulation Amount)가 필요해 최종적으로 종이에 인쇄되기 전 낭비되는 잉크가 발생하게 된다. 수집 한 데이터를 통해 총 사용한 잉크(Ink Used)와 Ink Usage($g/m^2$)을 Scatter Plot을 통해 나타내어 <Figure 2>에 나타내고 총 작업한 인쇄 매수(Product Qty)와 Ink Usage($g/m^2$)을 Scatter Plot을 통해 나타내어 <Figure 2>에 나타낸다.




<Figure 2> Ink Used (g) Scatter Plot

사용한 잉크량(Ink Used)과 Ink Usage($g/m^2$)의 관계를 나타낸 <Figure 2>에서 확인할 수 있듯이 사용한 잉크량이 많을수록 Ink Usage($g/m^2$)가 줄어드는 것을 확인할 수 있다. 위와 같은 경향을 보이는 이유는 사용되는 잉크인 Ink Used(g)가 늘어날수록 낭비되는 Minimum Circulation Amount(g) 양이 전체 양에서 차지하는 비율이 줄어들어 Ink Used(g)가 늘어날수록 Ink Usage($g/m^2$) 값은 줄어들 것이다.




<Figure 3> Product Quantity Scatter Plot

작업 매수(Product Quantity)와 Ink Usage($g/m^2$)의 관계를 나타낸 <Figure 3>에서 확인할 수 있듯이 작업한 매

수가 많을수록 Ink Usage(g/m$^2$)가 줄어드는 것을 확인할 수 있다. 작업 매수가 늘어날수록 3.1.(4)의 시험인쇄 과정 중 낭비되는 잉크량이 차지하는 비율이 줄어들게 된다. 마찬가지로 작업매수가 늘어날수록 Ink Usage(g/m$^2$)는 줄어들 것이다.

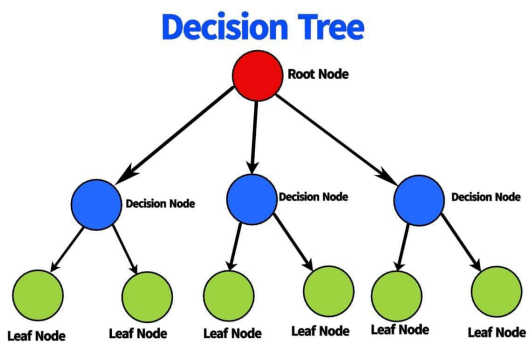
위 <Figure 2>와 <Figure 3>를 통해 확인할 수 있듯이 실제 작업현장에서는 작업에 사용한 총 잉크량과 작업매수와 비선형적 상호작용 효과가 있는 것을 확인할 수 있다. 그러므로 앞서 기존에 사용 중인 SPANKS식은 해당 특성을 반영하기 힘들고 작업자 직관에 계산되는 방식도 정확하게 필요량을 예측하기 힘들다. 그러므로 변수들 간의 비선형적인 특성을 반영할 수 있는 머신러닝 알고리즘이 요구된다.

## 3.4 머신러닝 모델

### 3.4.1 다중회귀분석 모델(Multiple Regression Analysis)

회귀분석은 독립변수(Independent Variable)와 종속변수(Dependent Variable)간의 상호관계(선형, 비선형 관계 등)를 분석하는 기법이다. 그 중 다중회귀분석[23](Multiple Regression Analysis)은 두개 이상의 독립변수들과 하나의 종속변수의 관계를 분석하는 기법이며 단순회귀 분석을 확장한 모델로 본 연구에서는 다른 머신러닝 모델을 평가할 Baseline 모델로 사용하였다. 종속변수 $Y$는 예측 잉크 필요량이다. 다중회귀식을 사용할 경우 모든 독립변수들을 한 번에 포함하여 분석할 수 있으며 특정 독립변수의 영향력을 알 수 있다. 다중 회귀 모델과 같은 경우 독립변수 값의 변화에 따른 종속변수 값의 변화가 일정해야 하는 독립변수와 종속변수 간의 선형성을 만족해야 한다.

### 3.4.2 의사 결정나무 모델(Decision Tree Model)



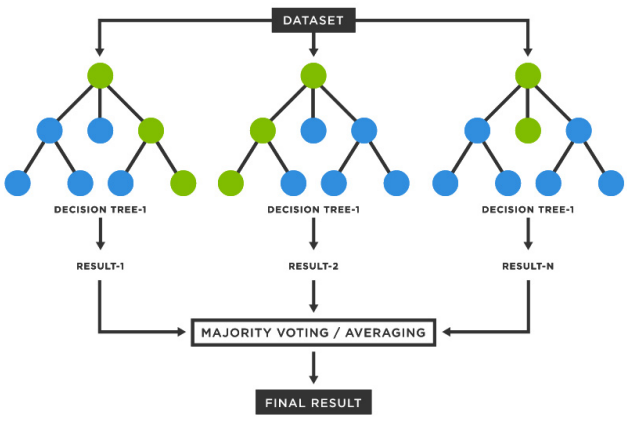

<Figure 4> Decision Tree Example

의사결정나무[21]는 여러 가지 규칙을 순차적으로 적용하면서 독립 변수 공간을 분할하는 분류 모형이다. 분류와 회귀 분석에 모두 사용될 수 있어서 CART(Classification And Regression Tree)라고도 한다. 의사 결정나무 모델의 분류법은 다음과 같다. 복수의 독립 변수 중 하나의 독립변수를 선택하고 기준 값(threshold)이라 정한다. 전체 학습 데이터(Root Node)를 해당 독립 변수의 값이 기준 값 보다 작은 데이터 그룹(Decision Node)과 해당 독립 변수의 값이 기준 값 보다 큰 데이터 그룹(Leaf Node)으로 나눈다. 각각의 자식 노드에 대해 앞에 단계를 반복하여 하위 자식 노드를 생성한다. 하지만 자식 노드에 한 가지 클래스의 데이터만 존재한다면 더 이상 자식 노드를 나누지 않고 중지한다. 이렇게 자식 노드 나누기를 연속적으로 적용하면 노드가 계속 증가하는 나무(tree)와 같은 형태로 표현할 수 있다. 본 연구에서 잉크 필요 예측량을 구하기 위해 사용한 의사결정 나무 형식은 회귀나무(Regression Trees)이며 출력변수가 연속형으로 나온다.

### 3.4.3 랜덤 포레스트 모델(Random Forest)

랜덤 포레스트(Random Forest) 알고리즘[22]은 다수의 의사 결정나무를 조성하여 예측을 진행하는 머신러닝 알고리즘이다.

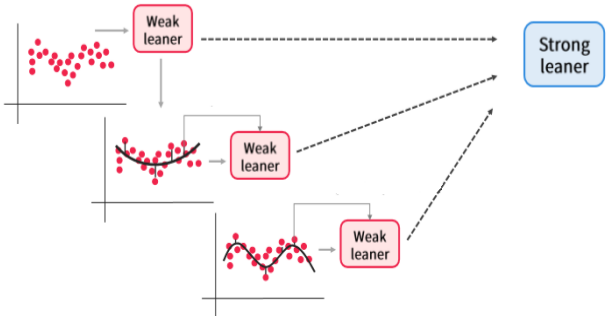
랜덤포레스트는 위 의사결정나무를 여러 개 파생시켜 각 의사결정 나무에서 나온 결과를 합하여 제공한다. 랜덤포레스트는 각 의사결정나무를 만들 때 랜덤으로 학습 데이터와 독립 변수를 선택하여 예측을 진행한다. 이 방법으로 만들어진 개별적인 의사결정나무는 정확도는 떨어질 수 있으나 모든 의사결정나무를 종합하여 예측을 수행하므로 정확도와 안정성이 높아진다는 장점을 지닌다. 즉, 랜덤포레스트는 무작위로 독립 변수를 N개 고르고, 데이터 또한 무작위로 선정하는 의사결정나무 알고리즘을 T개 만들어 다수결의 원칙으로 가장 많이 도출되는 값 또는 평균값을 예측값으로 사용한다.




<Figure 5> Random Forest Example

#### 3.4.4 GBM 모델(Gradient Boosting Machine)

GBM(Gradient Boosting Machine)[16]은 부스팅 앙상블 알고리즘이다. 부스팅이란 여러 개의 Weak Learner를 순차적으로 학습하면서 각 step에서 잘못 예측된 데이터에 대해 가중치를 부여해 오류를 개선해 나가는 학습 방식이다. GBM은 CART 기반의 다른 알고리즘과 마찬가지로 분류뿐만 아니라 회귀 문제에도 사용될 수 있다. GBM은 앞에서 학습한 모형의 틀린 예측 데이터들을 고쳐 나가는 방식을 사용하는데 의사결정 나무를 사용하고 예측 오류를 최소화해 나가는 과정에서 경사하강법(Gradient Descent)을 이용한다. 마지막에는 이 weak learner들이 결합하여 성능이 좋은 하나의 모델(Strong Learner)이 만들어진다.




<Figure 6> Gradient Boosting Machine Example

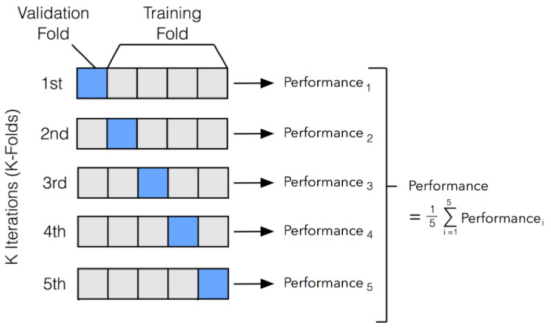
#### 3.4.5 XGBoost 모델

XGBoost[4]는 약한 예측 모델인 트리 모델을 결합하여 강한 예측 모델을 만드는 알고리즘으로서 부스팅 기법을 이용하여, 순차적으로 오류를 줄여나간다. 먼저 약한 분류기가 입력 데이터를 통해 학습되면 학습된 결과에서 나타나는 오차를 또 다른 약한 분류기에서 학습시킴으로써 순차적으로 발생하는 오차 분포를 모델이 학습하게 된다. XGBoost는 부스팅 기법을 통해 모델들을 통합할 때 중요도가 높은 트리 모델에 높은 가중치를 부여하는데 N 번째 모델이 가지는 가중치는 N-1 번째 모델의 오류에 따라서 결정된다. 학습을 위한 목적함수는 참값과 예측값 사이의 손실 함수와 모델의 복잡도를 나타내는 정규화 항으로 구성된다.

이러한 목적함수를 최적화 시키는 방향으로 모델이 학습하게 되며 목적함수를 최소화하는 가중치를 구하게 된다. 또한 XGBoost는 서른 개 이상의 하이퍼 파라미터(hyper parameter)를 지원하므로 다양한 상황에 따른 데이터에 대한 유연한 학습이 가능하며, CART[14] 방식을 이용해 생성된 트리모델과 리프(leaf)의 우위를 비교하여 입력 특징의 중요도를 확인할 수 있다.

### 3.5 데이터 모델링

#### 3.5.1 K-Fold Cross Validation




<Figure 7> K-Fold Method (K=5)

본 연구에서는 인쇄 포장공장 중 하나인 국내 업체(덕수산업)에서 발생한 데이터를 수집하여 활용한다. 데이터 셋은 1034개의 샘플로 구성되어 데이터 수가 적기 때문에 모델의 성능이 민감하게 변할 수 있는 단점이 있으므로 이를 보완하기 위하여 k-겹 교차 검증[3]을 사용하였다. 데이터를 동일한 크기를 가진 k개의 데이터 셋으로 분할하고, 첫 번째 데이터 셋은 검증 데이터로 사용하고 나머지 k-1개의 데이터 셋에 대해 훈련하여 검증 데이터 셋의 오차를 계산한다. 이 과정을 순차적으로 k번 반복하여 얻어지는 오차 값들을 평균하여 계산한다. 이 검증방법은 지나치게 높은 편향과 높은 분산으로 인해 발생하는 문제없이 검증을 진행할 수 있다는 장점이 있다. 본 연구에서는 k=5인 교차 검증을 사용하였다.

#### 3.5.2 모델 성능 지표

본 연구에서 각 모델 별 성능을 비교하기 위한 지표로 MAE, MSE, R-squared를 통해 비교한다. MAE는 모델의 예측값과 실제값의 차이(절댓값)이며 MAE가 높을수록 성능이 낮다. RMSE는 모델의 예측값과 실제 관측 값 차이의 제곱의 합의 루트 값이며 RMSE가 높을수록 성능이 낮다.

R-squared(결정 계수)는 실제 관측 값의 분산대비 예측값의 분산을 계산하여 데이터 예측의 정확도 성능을 측정하는 지표이다. 0~1까지 수로 나타내어지며 1에 가까울수록 100%의 설명력을 가진 모델이라고 평가하게 된다. R-squared는 SST(Total Sum of Squares)와 SSE(Explained Sum of Squares)를 통해 다음과 같이 나타낼 수 있다.

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum(y_i - \hat{y})^2}{\sum(y_i - \bar{y})^2}$$

Adjusted R-squared는 다변수 회귀분석에서 독립변수

가 유의하든, 유의하지 않든 독립변수의 수가 많아지면 결정계수가 높아지는 단점을 보완하기 위해 사용한다. N은 총 샘플의 수이며 P는 독립변수들의 개수이다.

$$Adjusted\ R^2 = 1 - \frac{(1-R^2)(N-1)}{N-P-1}$$

모델의 용이성을 확인할 수 있는 절대적인 R-squared 기준은 없다[24]. 하지만 동일한 문제 내에서 사용하는 모델들의 설명력 비교 하기위해 R-squared를 사용한다.

마지막 성능 지표로 작업자의 직관에 의해 제조했던 조색양에서 머신러닝 모델을 통해 예측했을 경우 절약되는 잉크량을 통해 머신러닝 모델들의 성능을 비교한다.

#### 3.5.3 하이퍼파라미터 선택

모델의 성능을 비교할 때는 특별한 조정 없이 <Table 2>에 나타낸 하이퍼파라미터 값을 넣어 학습한다. 위 모델 중 가장 성능이 좋은 모델을 골라내어 파이썬 OPTUNA[1] 라이브러리 알고리즘을 통해 하이퍼 파라미터를 최적화한다. Optuna는 하이퍼 파라미터 최적화 프레임워크로 사용자가 하이퍼 파라미터를 일정 범위로 정하고 이 중 최적화를 위해 설정된 측정 지표에 따라 하이퍼 파라미터를 지속해서 수정해 나가는 방식이다.

## 4. 결과 및 고찰

각 머신러닝 모델들의 비교 결과를 <Table 3>에 나타냈다. 다중회귀분석과 같이 R-squared가 Negative가 나오는 경우는 선형 회귀모델에서 각 독립변수의 평균값으로 예측하였을 경우보다 성능이 낮다는 것을 의미한다. 그러므로 선형 회귀분석으로는 잉크 필요량 예측 문제에 적합하지 않다[23].

CART 기반의 알고리즘인 의사 결정 나무부터 MAE, MSE, R-squared 모두 다중 선형회귀 모델 보다 좋은 결과를 보였다. 그 중 성능이 가장 좋은 모델은 XGBoost 모델로 나왔으며 MAE = 657, RMSE = 1087, R-squared = 0.70, Adj R-squared = 0.64이 나오며 해당 모델을 사용할 경우 기존 작업자가 예측하던 때에 비해 작업 하나당 2114 (g)의 잉크를 절약할 수 있는 결과가 나왔다. 가장 성능이 좋았던

<Table 2> Hyper Parameter of Machine Learning Models

| Model | Hyper Parameter Values |
|---|---|
| Multiple Regression | None |
| Decision Tree | random_state = 0, max_depth = 50 |
| Random Forest | n_estimators = 100, random_state = 0, min_sample_split = 10 |
| GBM | n_estimators = 100, random_state = 0, min_sample_split = 10 |
| XGBoost | max_depth = 50, n_estimators = 100, random_state = 0, colsample_bytree = 0.7, min_child_weight = 1 |

<Table 3> Machine Learning Model Performance Comparison

| Model | MAE | RMSE | R-squared | Adj. R-squared | Amount of ink savings |
|---|---|---|---|---|---|
| Multiple Regression | 1350 | 1974 | -0.08 | -0.16 | 2009(g) |
| Decision Tree | 1004 | 1695 | 0.46 | 0.42 | 1924(g) |
| Random Forest | 715 | 1201 | 0.64 | 0.56 | 2064(g) |
| GBM | 716 | 1164 | 0.63 | 0.55 | 2032(g) |
| XGBoost | 657 | 1087 | 0.70 | 0.64 | 2114(g) |

<Table 4> Optimizing XGBoost Model Hyper Parameter

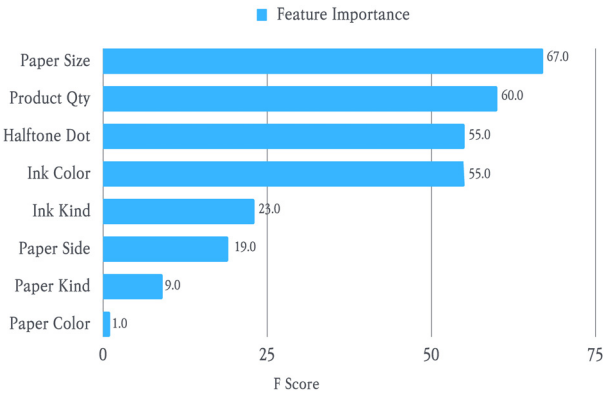
| Hyper Parameter | Description | Optimization Range | Optimization Value |
|---|---|---|---|
| max_depth | The maximum depth of a tree | [1~100] | 15 |
| colsample_bytree | the fraction of columns to be randomly samples for each tree | [0.5~1] | 1 |
| min_child_weight | Minimum sum of instance weight | [1~50] | 7 |
| n_estimators | Number of trees boosted | [30~10000] | 78 |

<Table 5> Hyper Parameter Optimized XGBoost Model

| Model | MAE | RMSE | R-squared | Adj. R-squared | Amount of ink savings |
|---|---|---|---|---|---|
| XGBoost (Optuna) | 642 | 1051 | 0.71 | 0.67 | 2134(g) |

XGBoost 모델의 하이퍼 파라미터 최적화를 위해 최적화할 파라미터 변수로는 <Table 4>에 나타냈다.

최적화된 파라미터 기반으로 XGBoost 모델에 적용한 결과 모델의 성능이 상승한 것을 <Table 5> 통해 확인할 수 있다.




<Figure 8> XGBoost Model Feature Importance

최종 XGBoost의 모델의 Feature Importance를 확인한 결과 <Figure 8>과 같이 Paper Size, Product Qty, Halftone Dot, Ink Color, Ink Kind, Paper Side, Paper Kind, Paper Color 순으로 변수 중요도가 높았다. Ink Color와 같은 명목형(Categorical) 변수의 중요성이 다른 연속형 변수들과 마찬가지로 비슷하게 높은 수치가 나온 만큼 잉크 필요 예측량 문제에서 명목형(Categorical) 변수들의 중요성을 알 수 있다. 또한 3.3장에서 확인한 작업매수와 잉크 총 사용량과 관련된 비선형적인 관계를 표현할 수 있는 모델이 필요하다. 그러므로 비선형적인 관계의 변수들과 명목형(Categorical) 변수들을 다뤄야 하는 패키징 프린팅의 잉크 필요량 예측 문제에서 다중선형회귀 모델은 적합하지 않으며 예측 결과도 R-squared가 음의 값이 나오는 것이 이를 뒷받침한다.

변수들 간의 비선형적인 관계와 명목형 변수의 데이터를 다룰 때 성능이 우수한 CART 기반의 의사결정 나무부터 R-squared가 양의 값을 가지는 것을 확인할 수 있었고 MAE, MSE 다중 선형회귀 모델보다 높았다. 이를 통해 잉크량 예측 문제에 있어 CART 기반의 모델이 적합하다는 것을 확인 수 있었다. 또한 CART 기반의 모델인 의사 결정 나무, 랜덤 포레스트, GBM, XGBoost 중 XGBoost의 성능이 가장 뛰어난 것을 확인할 수 있었다.

가장 높은 성능을 보였던 XGBoost의 R-squared 값이 0.70인 것은 해당 모델이 70%의 데이터의 변동성에 대해 설명이 용이하지만 나머지 30%에서는 설명이 불가능하다는 것을 의미한다. 머신러닝 모델을 사용할 경우 작업자의 직관으로 예측하는 때 보다 작업 하나 당 2114(g)의 잉크를 절약할 수 있었지만 아직 잉크 필요량의 정밀한 예측을 위해서는 머신러닝 모델에 대한 연구가 더 필요한 실정이다.

## 5. 요약 및 향후 연구

본 연구를 통해 사람의 직관에 의해 진행되던 패키징 프린팅 작업에서 머신러닝 모델 중 명목형(Categorical) 변수와 비선형적인 데이터를 다룰 때 우수한 CART 기반의 모델들이 적합하다는 사실을 알 수 있었다. CART 기반의 모델 중 XGBoost가 다른 모델들에 비해 성능이 우수했으며 사람에 직관에 의해 제조될 때 보다 매 작업마다 평균 2134(g)의 잉크를 절약할 수 있었다.

향후 머신러닝의 예측 성능을 향상시키기 위해 종이의 물성치와 관련된 새로운 명목형(Categorical) 학습변수로 종이의 코팅 유무와 코팅재질 변수를 수집하여 활용할 계획이다. 또한 향후 연구에서는 예측량이 필요량보다 낮아 작업이 중단되는 경우를 방지하고자 XGBoost의 목적함수(Objective Function)를 Customizing하여 잔차(Residual)가 남는 모델을 개발할 것이다.

본 논문의 결과를 바탕으로 향후 잉크 필요량 예측 연구에서는 머신러닝 알고리즘 중 CART 기반의 회귀 모델을 사용할 것을 제안한다.

## Acknowledgement

This research was a part of the project titled 'forest science-technology R&D program (2021383A00-2223-0101)', funded by the Korea Forestry Promotion Institute (Korea National Arboretum), Korea.

This study has been funded by the following project of KITECH (Korea Institute of Industrial Technology), which is “Development of AI-based packaging manufacturing cost estimation system”.

## References

[1] Akiba, T., S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data mining*, 2019.

[2] Ash, A. and Shwartz, M., $R^2$: A Useful Measure of Model Performance when Predicting a Dichotomous

Outcome, *Statistics in Medicine*, 1999, Vol. 18, No. 4, pp. 375-384.

[3] Braga-Neto, U.M. and Dougherty, E.R., Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 2004. Vol. 20, No. 3, pp. 374-380.

[4] Chen, T. and Guestrin, C., Xgboost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM Sigkdd International Conference on Knowledge Discovery and Data Mining*, 2016.

[5] Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., and Chen, K., Xgboost: Extreme Gradient Boosting, *R package version 0.4-2*, 2015, Vol. 1, , No. 4, pp. 1-4.

[6] Dyer, D., Dalzell, F., and Olegario, R., *Rising Tide: Lessons from 165 Years of Brand Building at Procter & Gamble*, Harvard Business School Press Boston, MA, 2004.

[7] Friede, G., Busch, T., and Bassen, A., ESG and Financial Performance: Aggregated Evidence from More than 2000 Empirical Studies, *Journal of Sustainable Finance & Investment*, 2015, Vol. 5, No. 4, pp. 210-233.

[8] Friedman, J.H., Greedy Function Approximation: A Gradient Boosting Machine, *Annals of Statistics*, 2001, pp. 1189-1232.

[9] Gkerekos, C., Lazakis, I., and Theotokatos, G., Machine Learning Models for Predicting Ship Main Engine Fuel Oil Consumption: A Comparative Study, *Ocean Engineering*, 2019, Vol. 188, p. 106282.

[10] Gladysz, B., Krystosiak, K., Ejsmont, K., Kluczek, A., and Buczacki, A., Sustainable Printing 4.0—Insights from a Polish Survey, *Sustainability*, 2021, Vol. 13, No. 19, p. 10916.

[11] Jordan, M.I. and Mitchell, T.M., Machine Learning: Trends, Perspectives, and Prospects, *Science*, 2015, Vol. 349, No. 6245, pp. 255-260.

[12] Jung, H. and Kim, J.-W., A Machine Learning Approach for Mechanical Motor Fault Diagnosis, *Journal of Korean Society of Industrial and Systems Engineering*, 2017, Vol. 40, No. 1, pp. 57-64.

[13] Kaliya, G., *Estimation of Ink Consumption Used for Printing Process*, 2006.

[14] Lewis, R.J., An Introduction to Classification and Regression Tree (CART) Analysis, in *Annual meeting of the society for academic emergency medicine in San Francisco*, California, Citeseer, 2000.

[15] Murthy, S.K., Automatic Construction of Decision Trees from Data: A Multi-disciplinary Survey, *Data Mining and Knowledge Discovery*, 1998, Vol. 2, No. 4, pp. 345-389.

[16] Natekin, A. and Knoll, A., Gradient Boosting Machines, a Tutorial, *Frontiers in Neurorobotics*, 2013, Vol. 7, p. 21.

[17] Neter, J., Kutner, M.H., Nachtsheim, C.J., and Wasserman, W., *Applied Linear Statistical Models*, 1996.

[18] Razi, M.A. and Athappilly, K., A Comparative Predictive Analysis of Neural Networks (NNs), Nonlinear Regression and Classification and Regression Tree (CART) Models, *Expert Systems with Applications*, 2005, Vol. 29, No. 1, pp. 65-74.

[19] Ruggles, P.K., *Printing Estimating: Costing and Pricing Print and Digital Media,* PIA/GATFPress, 2008.

[20] Shehadeh, A., Alshboul, O., Al Mamlook, R.E., and Hamedat, O., Machine Learning Models for Predicting the Residual Value of Heavy Construction Equipment: An Evaluation of Modified Decision Tree, LightGBM, and XGBoost Regression, *Automation in Construction*, 2021, Vol. 129, p. 103827.

[21] Song, Y.-Y. and Ying, L., Decision Tree Methods: Applications for Classification and Prediction, *Shanghai Archives of Psychiatry*, 2015, Vol. 27, No. 2, p. 130.

[22] Svetnik, V., Liaw, A., Tong, C., Culberson, J.C., Sheridan, R.P., and Feuston, B.P., Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling, *Journal of Chemical Information and Computer Sciences*, 2003, Vol. 43, No. 6, pp. 1947-1958.

[23] Uyanık, G.K. and Güler, N., A Study on Multiple Linear Regression Analysis, *Procedia-Social and Behavioral Sciences*, 2013, Vol. 106, pp. 234-240.

**ORCID**

Se Wook Kwon | https://orcid.org/0000-0003-4556-8292
Young Joo Hyun | https://orcid.org/0000-0002-3119-275X
Hyun Chul Tae | https://orcid.org/0000-0002-2277-0722