한국빅데이터학회지
제4권 제2호, 2019, pp. 115-126

무역수출 라이브지수를 활용한 중소수출기업 발굴 연구

A Study on Detection of Small Export Companies Utilizing Trade Exports Live Index

김희천[1]·임춘성[2†]·성주원[3]

연세대학교 일반대학원 융합기술경영공학과[1], 연세대학교 산업공학과[2], ㈜미소정보기술[3]

요 약

무역수출 분야에서 수출 지수에 관한 논의는 수차례 있었으나 객관적 지표로 설명할 수 있는 명확한 무역수출 지수는 없다. 한국무역협회(KITA), 대한무역투자진흥공사(KOTRA) 등에서 지표를 만들고자 하는 시도를 하고 있으나 수출기업의 역량을 표현할 수 있는 방법에 대하여 현재 계속 고민 중이다. 이에 본 연구는 기업의 규모, 신용도와 같은 공시지표와 거래고객수, 거래횟수, 상품개수, 거래량, 거래기간 등의 활동지표를 feature로 설정하여 인공지능 학습 데이터 셋을 구축하고, 딥러닝 알고리즘에서 Lightgbm을 이용하여 수출 가능 기업에 대한 분류 모델을 제시한다. 또한 기업이 속한 산업 군집 분류 모델로 Graph Neural Network을 사용하여 기업간, 품목간, 사업군에서의 수출 가능 역량을 표현하는 수출 Live지수를 산출하였으며 이는 지수를 산출하는 현재로부터 기업의 과거 활동을 포함함으로써 객관성을 확보하였다.

■ 중심어: 수출지수, 종합 라이브지수, 중소기업, 구매확인서, 내국신용장

Abstract

There have been many discussions on export indices in trade exports, but there is no definite trade export index which can be explained by objective indicators. Korea International Trade Association (KITA), Korea Trade-Investment Promotion Agency (KOTRA), etc., but we are currently in the process of thinking about ways to express the capabilities of exporting companies. In this study, we constructed the AI data sets by setting the activity indicators such as the size of the company and the credit score, the number of transaction customers, the number of transactions, the number of items, the transaction volume, and the transaction period as features, Lightgbm. Using the Graph Neural Network as an industrial cluster classification model, the export live index which expresses the exportable capacity among companies, items, and business groups was calculated. This includes the past activity of the company from the current calculating index Objectivity.

■ Keywords: Export KPI, Comprehensive Live Index, Small or Medium Sized Business, Approval of Purchase, Local L/C

2019년 11월 29일 접수; 2019년 12월 10일 수정본 접수; 2019년 12월 31일 게재 확정.
† 교신저자 (leem@yonsei.ac.kr)

Ⅰ. 서 론

수출을 목표로 하고 있는 많은 기업들이 수출 역량을 판단할 수 있는 근거가 부족한 현 시점에서 KOTRA(대한무역투자진흥공사)에서 무역박람회, 수출대전 등을 통하여 기업의 역량 진단 및 상담, 바이어 매칭 등을 추진하고 있다. 국내에서 수기화(수출기업화) 사업을 목표로 하고 있는 KITA(한국무역협회)와 같은 기관에서도 해외 바이어 매칭을 위해 해외 inquery(구매문의)를 통해 국내 기업을 조사하고 평가하여 수출을 추천하는 프로세스를 가지고 있다. KOTRA는 2018년 ISP 사업을 시작으로 2019년 빅데이터 플랫폼 구축을 위한 사업을 시작하였으며 수출역량 자가진단 서비스를 갖고 있다.

이와 같은 활동들에서 선행되어야 할 전제는 수출을 원하는 기업들이 어느 정도의 역량을 가지고 있는지 객관적인 지수로 판단할 수 있는 근거가 필요하다는 것이다. 이 지수는 첫 번째 기업의 수출역량을 판단하고 수출을 원하는 기업의 진단 척도로 사용될 수 있으며, 두 번째 기업의 생산-공급 능력과 수준을 판단하여 적절한 수준의 기업끼리의 매칭이 가능하다. 세 번째 기업 간 매칭을 추진하는 기관들의 빠른 판단 근거로 사용될 수 있다. 해외 바이어 문의에 국내 기업을 빠르게 찾고 역량을 판단해야 하는 무역협회와 같은 입장에서 일차적으로 수출역량을 제시해줄 수 있는 지표가 있다는 것은 업무 효율성을 획기적으로 높일 수 있기 때문이다.

본 연구는 이러한 효용성을 위하여 무역수출 라이브지수 (Export Live Index)를 연구하여 단순 지표가 아닌 역동적인 지표로 활용되고자 한다.

Ⅱ. 이론적 배경 및 관련 연구

2.1 무역 데이터 활용

현재 무역수출 업무에 있어서 전자업무를 대행하고 있는 KTNET(한국무역정보통신)은 eTradeHub라는 플랫폼 활용하여 구매확인서와 내국신용장을 발급하는 서비스를 이행하고 있다.

수출은 크게 직접수출과 간접 수출로 구분할 수 있는데 직접수출은 직접 해외 바이어에게 물건을 공급하는 수출기업의 활동을 뜻하며, 간접수출은 이 직접수출 물품을 국내 혹은 국외에서 생산하기 위한 제품, 재료를 공급하는 활동을 수출 활동으로 인정받는 것을 뜻한다. 즉, A기업이 C라는 제품을 생산하여 수출하는 데에 B기업의 D라는 제품이 부속품으로 들어간다면, A기업은 직접 수출기업, C는 수출품목, B는 간접 수출기업, D는 간접 수출품목이 된다.

이에 대해서 A기업은 B기업에게 구매했다는 확인서를 발급해 줄 수 있게 되는데 이때, B기업은 수출에 기여한 간접 수출 기업으로서 수출기업이 가질 수 있는 각종 세금 혜택 등을 받을 수 있게 된다. 구매확인서와 마찬가지로 내국신용장(Local L/C)은 수출업무에서 수출업자의 거래은행이 대금 지급을 보증해주는 해외 발급 신용장을 담보로 국내에서 개설해주는 제 2의 신용장이다.

본 연구는 KTNET으로부터 비식별화된 3년(2016-2018) 분량의 구매확인서, 내국신용장 데이터를 열람할 수 있는 권한을 한시적으로 받아서 내부 보안정책을 준수하고 현장에서 분석 및 결과를 도출하는 과정을 거쳤으며 이에 약 4000만 건의 거래 건수에 대한 약 13만 건의 기업을 데이터로 분석하였다.

2.2 추가 신용등급 데이터 확보

기업의 신용정보를 제공하는 기관인 CRETOP

의 유료서비스를 이용하여 약 13만 건의 기업에 대한 신용등급과 현금 흐름 등급을 같은 해인 2016-2018년 3년 치를 추가 수집하였다.

이는 구매확인서와 내국신용장은 기업의 수출 역량을 판단하는 것에 주요한 자료원으로 사용될 수 있으나 신용등급과 현금흐름 등급은 수출 외 기업 활동에 대한 총체적 지표로서 보완자료로 활용될 수 있기 때문이다. 또한 구매확인서를 발급받고 있는 기업들은 이미 간접 수출기업으로 인정받은 기업이므로 앞으로 수출을 지향하는 기업들이 수출 기업화하기 위하여 수출지수를 이용할 경우 신용등급과 현금흐름등급이 1차 대입값으로 활용될 수 있다.

신용등급과 현금흐름등급 외 다른 지표들이 연구에 도움을 줄 수 있음을 판단하였으나 기업의 민감정보, 혹은 영업비밀에 해당하는 자료들이 통계적으로라도 이용되는 것이 위험할 수 있으므로 두 가지 등급만을 활용하였다.

2.3 데이터 정제

구매확인서와 내국신용장이 가지고 있는 정보는 구매기업과 공급기업, 해당 거래에 이동된 품목, 거래량, 거래금액, 거래날짜이다. 이로써 거래날짜를 통하여 얻을 수 있는 통계적 지표를 다양하게 산출하였으며 이는 다음과 같다.

거래량에서 얻을 수 있는 기업의 생산 능력 및 공급 규모, 거래금액에서 얻을 수 있는 기업의 실적 규모가 그 첫 번째이다. 두 번째는 거래날짜를 통해서 알 수 있는 정보이며 여러 가지 학습에 필요한 변수를 준비할 수 있다.

거래날짜로부터 해당 기업은 얼마나 오랫동안 거래를 지속해 왔는지 파악할 수 있다. 이는 기업이 얼마나 꾸준히 간접 수출 활동을 이어왔는지에 중요한 지표로 활용된다. 또한 연중 1, 2회로 단발성 거래인지 그 이상 고르게 분포된 거래상태를 나타내는지를 고려해볼 때 기업의 안정성을 대략적으로 파악할 수 있다. 마지막으로 각 거래에 대한 거래금액 편차를 대입함으로 인하여 현금흐름도 대략적으로 파악이 가능하다.

마지막으로 거래 대상 기업이 다양한지 혹은 품목이 다양한지, 거래 빈도가 잦은지에 대한 값을 도출하여 기업이 해당 수출 활동에 대한 분산적 risk taking이 가능한 기업인지에 대한 척도를 판단할 데이터로 확보하였다.

2.4 LightGBM

그래디언트 부스팅 결정 트리(GBDT : Gradient Boosting Decision Tree)는 자주 이용되는 기계학습 알고리즘으로서 효율성이나 정확도 측면에서 매우 우수하고 해석 가능성이 높다는 점에서 널리 사용되고 있다. 그러나 최근 빅데이터가 학습 데이터로 활용되기 시작하면서 GBDT가 정확도와 효율성 간의 트레이드 오프라는 숙제를 떠안게 되었다. GBDT의 원리는 각 변수마다 가능한 모든 분할점에 대해서 정보 획득을 평가하기 위해 데이터 개체를 모두 훑어야 하는 구조이다. 이에 따라 계산식의 복잡성이 변수 개수와 개체 수에 비례하게 마련이다. 그러므로 빅데이터 처리 시 매우 시간 소모적인 측면이 있다.

이러한 문제 해결을 위해 기울기 기반 단측 표본 추출(GOSS : Gradient-based One-side Sampling)과 배타적 변수 묶음(EFB : Exclusive Feature Bundling)이라는 두 가지 새로운 개념을 도입하여 기울기가 큰 데이터 개체가 정보 획득 계산에서 더 중요한 역할을 한다는 것을 전제로 하여 훨씬 작은 크기의 데이터에서 정보를 정확히 추정해 낼 수 있도록 하였다. 또한 EFB를 활용하여 변수 개수를 줄이는 상호 배타적 변수 즉, 0이 아닌 값이 동시에 발생하지 않는 변수들을 묶어 GBDT의 분할점 결정에 대한 정확도를 크게 해치지 않는 범위 내에서 변수 개수를 효과적으로 줄이는 것에 기여했다.

이러한 GOSS와 EFB를 적용하여 GBDT를 새

로 정의한 것을 LightGBM이라 하며 GBDT의 학습 훈련 시간을 최대 20배가량 줄이면서 정확도를 비슷한 수준으로 유지하는 모델이다.

Ⅲ. 연구 설계 및 연구 방법

본 연구는 python을 활용한 LightGBM 오픈 소스를 활용하여 변수 설정에 대한 파라미터를 조정하여 입력값에 대한 probability를 0-1사이 값으로 산출하여 1에 가까울수록 무역수출 지수가 높은 것으로, 0에 가까울수록 무역수출 지수가 낮은 것으로 판별하였다.

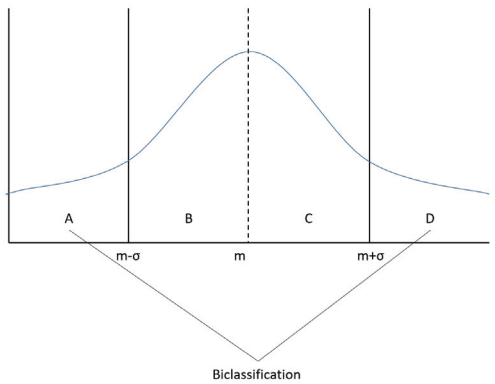
이에 따른 상위 모형은 데이터를 준비하고 탐색, 모델링과 모델평가의 순서로 모형이 만들어졌으며 세부 모형은 데이터 확보, 데이터 정제, 학습, 무역수출 지수 확보에 대한 분석 모형을 설계하였고 연구 설계 모형은 <그림 1>과 같다.

3.1 LightGBM 학습 데이터

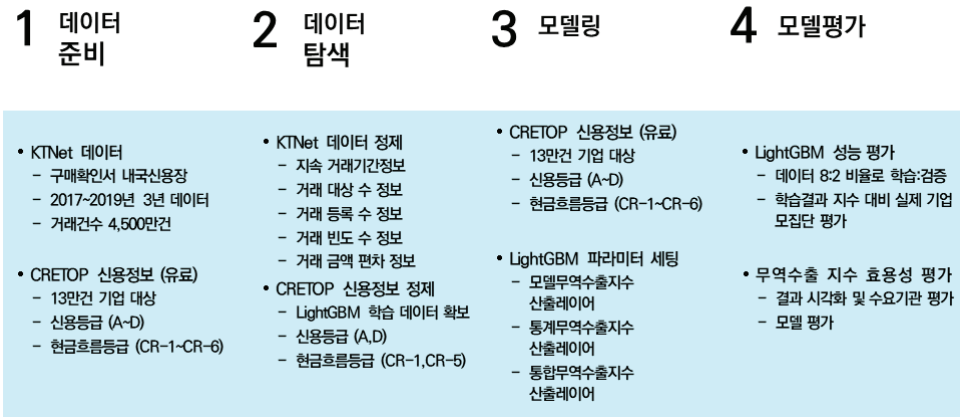
LightGBM모델이 빅데이터의 변수를 줄이는 것에 기여한다는 점에서 학습에 활용할 데이터도 multi-classification이 아닌 bi-classification<그림 2>을 활용한다. 즉 데이터의 변수분포를 정규분포로 나타내었을 때, 신용등급을 나타내는 A, B, C, D등급 중에 대부분을 차지하는 B, C등급은 데이터 모수도 많고 유사성도 높기 때문에 변별력이 매우 떨어지는 데이터이다.

이 부분을 제외하고 양극단의 데이터를 취하여 학습 데이터로 활용함으로써 모델의 성능을 높이고 정확도를 향상하는 효과를 얻어 내었다.

신용등급을 나타내는 양극단의 데이터인 A, D등급과 마찬가지로 현금흐름을 나타내는 CR1(매우 양호), CR5(보통 이하) 데이터로 양극단의 데이터로 추출하고 CR2, CR3, CR4는 학습에서 제외하였다. A등급과 CR1 등급을 probability 값의 1로 세팅, D등급과 CR5 등급을 0으로 세팅하여



〈그림 2〉 학습 데이터 bi-classification 영역



〈그림 1〉 연구 설계 모형

결과값이 0-1사이의 값으로 분류되도록 학습하였다.

(실제로 현금흐름등급은 CR1부터 CR6까지 6개의 등급으로 나누어지나 (판정제외, 판정보류는 논외) 실제로 CR6 등급을 받는 기업이 매우 극소수이기 때문에 CR5까지 5단계로 나누어 설정하였다.)

bi-classification으로 선별한 양극단의 값들을 선별하여 학습하여 결과를 도출하는 모델을 전체 기업에 적용하였을 때 각 기업이 갖는 0에서 1사이의 값에 분할점을 두어 일정 수준 이상이면 직수출 가능 기업, 일정 수준 이하면 아직 많은 준비가 필요한 기업, 그 사이의 기업들을 수출 컨설팅을 통해 보완하면 수출이 가능한 잠재 기업으로 분류하여 연구로 활용할 수 있도록 결과값이 도출된다.

3.2 데이터 모델링

무역수출지수 산출 레이어 도식은 <그림 3>과 같다.

모델무역수출지수 산출레이어의 학습에 사용되는 손실함수는 로지스틱 손실함수(logloss)로 정해질 수 있으며 log(odds)의 개념이 적용될 수 있다. 여기서 손실함수인 로지스틱 손실함수(logloss)는 다음과 같은 수식1로 정해질 수 있다.

[수식1]

$$\sum_{i=1}^{n} y_{i\times}\log_e(p)+(1-y_i)\times\log_e(1-p)$$

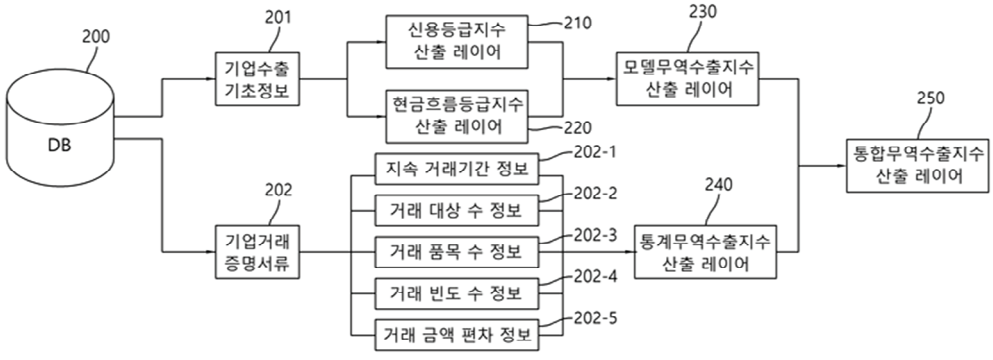
이때, n은 전체 데이터의 개수, i는 전체 데이터 중 특정 데이터 각각에 대응되는 수이며, 는 i번째 특정 데이터의 관측값이고, p는 i번째 특정 데이터의 예측값을 의미한다.

또한, 모델무역수출지수 산출 레이어로 하여금 의사결정트리의 생성을 위한 최초기준값 및 최초 예측값을 생성하도록 할 수 있는데, 이때 최초기준값은 log(odds)값으로서 전체 기업 각각의 기업 수출기초정보 중 첫번째 학습데이터에 대하여 상기 로지스틱 손실함수를 사용하여 산출되는 손실값의 합이 최소화될 수 있는 특정값을 의미하며, 결과적으로 아래의 수식으로 구해질 수 있다.

[수식2]

$$\log_e(odds)=\log_e\frac{(\text{성공확률})}{(\text{실패확률})}$$

이때, 성공확률은 실제 적용 시 수출기업의 비율에 대응될 수 있고, 실패확률은 수출기업이 아닌 기업의 비율에 대응될 수 있다. 예를 들어, 전



〈그림 3〉 무역수출지수 산출 레이어 도면

체 기업이 3개의 기업으로서, 2개의 기업이 수출기업이고 1개의 기업이 수출기업이 아닌 기업이라면, 최초기준값은 0.69가 될 수 있다.

또한, 최초예측값은 확률값으로서, log(odds)값인 최초기준값을 사용하여 아래의 [수식3]를 사용하여 구해질 수 있다.

[수식3]

$$p = \frac{e^{\log_e(odds)}}{1+e^{\log_e(odds)}} = \frac{odds}{(1+odds)}$$

따라서, 위의 예시와 같이 2개의 수출기업과 1개의 수출기업이 아닌 기업에 대해서 산출된 최초기준값이 0.69인 경우, 최초예측값은 0.67이 산출될 수 있다.

전체 기업 각각에 대하여 수출 여부와 최초예측값의 차이인 잔여 오차값을 구할 수 있는데, 위의 예시와 같이 3개의 기업에 대하여 최초예측값이 0.67로 구해진 상태라면 수출기업은 1, 수출기업이 아닌 기업은 0이므로 2개의 수출기업 각각의 잔여 오차값은 1 - 0.67 = 0.33, 1개의 수출기업이 아닌 기업의 잔여 오차값은 0 - 0.67 = -0.67이 될 수 있다.

모델무역수출지수 산출 레이어로 하여금 의사결정트리를 생성하도록 할 수 있는데, 전체 기업 각각의 기업수출기초정보 중 첫 번째 학습 데이터로서 입력받아, 분할 기준에 따라 a개의 부분집합인 학습 데이터 그룹으로 분할 할 수 있다. 이때, 상기 분할 기준은, 의사결정트리의 하위노드를 생성함에 있어 상위노드에 포함된 데이터를 어떻게 하위노드 각각으로 분할 할 것인지에 대한 기준으로서, 분류 트리의 경우 분할 전후의 데이터의 불순도를 비교하여 해당 불순도가 최대로 감소하도록 데이터를 분할할 수 있으며, 불순도의 계산 지표로는 지니 불순도, 정보 이득 개념에 기초한 엔트로피(불확실성) 등이 있다. 예를 들어, 분할 전의 데이터 전체의 엔트로피가 0.95이고, 분할 후 각각의 엔트로피를 합한 값이 0.75인 경우, 엔트로피가 분할 후에 감소하였으므로, 불확실성이 감소하고 정보 이득이 증가한 것으로 판단하여 해당 분할 기준에 따라 상위노드에 포함된 데이터를 분할 하고 그 각각에 대응되는 하위노드를 각각 생성할 수 있다.

모델무역수출지수 산출 레이어로 하여금 상술한 바와 같은 데이터의 분할 및 분할된 데이터 각각에 대응되는 하위노드의 생성을 소정의 트리 생성기준을 만족할 때까지 반복하여 k 깊이를 가지는 제1 트리를 생성하도록 할 수 있는데, 이때 상기 트리생성기준은 사용자에 의해 설정된 트리의 가지 및 마디 중 적어도 하나에 대한 최대 생성 가능 수치에 따라 결정될 수 있다.

제1 트리의 최종노드들이 생성되면, 프로세서는 모델무역수출지수 산출 레이어로 하여금 최종노드 각각에 대응되는 노드 대표값을 산출하도록 할 수 있다. 이때, 상기 노드 대표값 각각은, 상기 노드에 포함된 학습 데이터 그룹 각각의 잔여 오차값 및 이전예측값을 참조하여 아래와 같은 [수식4]에 의하여 산출될 수 있다.

[수식4]

$$\frac{\sum Residual_i}{\sum previous(p_i) \times 1 - previous(p_i)}$$

이때, i는 전체 데이터 중 특정 데이터 각각에 대응되는 수이고, 는 특정 데이터 각각에 대하여 예측된 예측값을 의미하며, 는 제1 트리의 생성 시에는 최초예측값이고 2 이상인 n에 대하여 제n 트리의 생성 시에는 제1 트리 혹은 제 n-1 트리까지에 의해 업데이트된 모델에 의해 예측된 이전예측값을 의미한다.

예를 들어, 수출기업 2개 및 수출기업이 아닌 기업 1개의 기업에 대하여 제1 트리의 특정 최종노드에 수출기업 1개, 수출기업이 아닌 기업 1개가 포함되었고, 각각의 기업의 잔여 오차값이 하

나는 0.33, 하나는 -0.67인 경우, 해당 최종노드의 노드 대푯값은 $\frac{0.33+(-0.67)}{(0.67\times(1-0.67))+(0.67\times(1-0.67)}$ = -0.77의 값이 산출될 수 있다.

최초기준값, 노드 대표예측값 및 기설정된 학습률을 반영하여 모델무역수출지수 산출 레이어의 모델이 업데이트되도록 할 수 있는데, 이는 최초기준값에 제1 트리에 의하여 산출된 노드 대표값을 기설정된 학습률을 반영하여 합하는 것일 수 있다. 따라서, 업데이트된 예측값은 최초기준값+(학습률*노드대표값)의 방법으로 산출된 log(odds)값을 사용하여 계산될 수 있는데, 기설정된 학습률이 0.8이고, 위의 예시에서와 같이 노드대표값이 -0.77인 최종노드에 해당되는 수출기업 1개와 수출기업이 아닌 기업 1개 각각의 log(odds)값은 0.67+(0.8*(-0.77)) = 0.05가 될 수 있고, 최종 예측값인 확률값은 상기 수식 4에 의하여 p = 0.5의 값이 될 수 있다.

업데이트된 모델무역수출지수 산출 레이어에 의하여 산출된 전체 기업 각각의 업데이트된 예측값 및 수출 여부 정보를 참조하여 잔여 오차값을 새롭게 각각 계산하고, 새로운 잔여 오차값을 사용하여 새로운 트리를 만드는 과정을 기설정된 트리생성조건에 만족될 때까지 반복하여 제 2 내지 제 m 개의 트리를 생성할 수 있으며, 각각의 트리가 생성될 때마다 모델무역수출지수 산출 레이어의 모델이 업데이트될 수 있다. 이때, 상기 트리생성조건에는 사용자에 의해 설정된 최대 트리 개수가 포함되어 있으며, 상기 제 m 트리가 최대 트리 개수에 해당되는 트리인 경우 학습이 종료될 수 있다.

그리고, 상술한 바와 같은 모델무역수출지수 산출 레이어의 학습 방법은 신용등급지수 산출 레이어 및 현금흐름 등급지수 산출 레이어에 대해서도 동일하게 적용하되, 신용등급지수 산출 레이어의 학습의 경우에는 제1 학습 데이터 대신 제2 학습 데이터가 신용등급지수 산출 레이어에 입력되도록 하고, 전체 기업 각각에 대한 수출 여부에 대한 정보 대신 상기 전체 기업 각각에 대한 신용등급에 대한 정보를 참조하여 최초기준값, 최초예측값 및 잔여 오차값을 산출하도록 하는 것을 특징으로 한다.

현금흐름 등급지수 산출 레이어의 학습의 경우에는 제1 학습 데이터를 대신하여 제3 학습 데이터가 현금흐름 등급지수 산출 레이어에 입력되도록 하고, 전체 기업 각각에 대한 수출 여부에 대한 정보 대신 전체 기업 각각에 대한 현금흐름 등급에 대한 정보를 참조하여 최초기준값, 최초예측값 및 잔여 오차값을 산출하도록 할 수 있다.

3.3 데이터 학습 순서

기울기 기반 단측 표본추출(GOSS)은 데이터의 분할 시 정보 획득 정도를 평가하기 위하여 모든 분할 가능 경우를 판단해야 함으로써 발생하는 시간, 노력을 줄이기 위해 판단 대상인 전체 데이터 각각의 기울기값을 참조한다. 기울기 값이 큰 상위데이터를 a라 할 때 a*100%를 선택하고 나머지 데이터를 b라 할 때 b*100%의 데이터를 무작위 표본 추출하여 가중치 ((1-a)/b)를 반영하여 증폭시킴으로써 보완된 데이터를 사용하는 방법으로 활용하였다.

배타적 변수 묶음(EFB)은 데이터 각각의 특성 및 특성값 각각에 대하여 상호 충돌이 일어나지 않거나 소정의 충돌비율 이하의 특성들을 하나의 특성으로 묶음(bundling) 처리함으로써 더 적은 수의 특성을 사용하도록 하는 방법으로 활용하였다. 이러한 방법을 사용하여 학습에 필요한 시간을 단축하고 더욱 효율적인 학습이 이루어지도록 설계하였다.

3.3.1 알고리즘 학습 프로세스

LightGBM에서 세팅하는 변수는 기본적으로 신용등급이 높고 현금흐름 등급이 높은 기업들이 무역수출 가능성이 높은 기업이라는 전제에서 출

발한다. 따라서 기업수출 기초정보 중 신용등급 지수와 현금흐름 등급을 제 1 부분 정보로 간주하고, 제 2 부분 정보를 수출 여부 정보로 생각하였다.

첫 번째 프로세스는 모델무역수출지수 산출 레이어가 전체 기업 각각에 대한 기업수출 기초정보 중 제 1부분 정보를 제 1 학습 데이터로 입력받아, 전체 기업 각각에 대한 현금흐름 등급 정보를 참조하여 최초기준값 및 최초 예측값을 산출한다.

산출된 최초예측값 및 전체 기업 각각에 대한 기업수출 기초정보 중 제 2 부분의 정보인 수출 여부 정보를 참조하여 전체 기업 각각에 대한 잔여오차값을 산출한다.

제 1 학습 데이터에 대응되는 하나의 root node를 생성하고, 제 1 학습 데이터를 분할기준에 따라 a개의 1-1, 1-2, ..., 1_a 학습데이터 그룹으로 분할하여, 분할된 학습데이터 그룹 각각에 대응되는 1_1, 1_2, .., 1_a 하위노드를 각각 생성한다.

생성된 하위노드 중 특정 하위노드에 대하여 그에 대응되는 특정 학습 데이터 그룹을 위 분할기준에 따라 2_1, 2_2, ..., 2_b 학습 데이터 그룹으로 분할하여, 분할된 학습 데이터 그룹 각각에 대응되는 2_1, 2_2, ..., 2_b 하위노드를 각각 생성하는 하위노드생성과정을 트리생성조건에 해당되는 횟수만큼 반복한다.

이 하위노드생성과정을 반복하여 k_1, k_2, ..., k_c 최종노드를 포함하는 k 깊이의 제 1 트리를 생성한다.

k_1, k_2, ..., k_c 최종노드 각각에 대하여 전체 기업 각각에 대한 잔여 오차값 및 이전예측값을 참조하여 노드 대표예측값을 각각 산출한다.

상기 노드 대표예측값 및 기설정된 학습률을 참조하여 모델무역수출지수 산출 레이어의 모델을 업데이트한다.

업데이트된 모델무역수출지수 산출 레이어에 의하여 예측되는 전체 기업 각각의 업데이트된 예측값 및 전체 기업 각각의 기업수출 기초정보 중 제 2 부분인 수출 여부에 대한 정보를 참조하여 새로운 트리를 생성하는 과정을 이 트리생성조건에 해당되는 횟수만큼 반복한다.

3.3.2 알고리즘 평가 및 반복 학습

사실상 학습 프로세스에 대해서 전부 알지 않아도 무관하다. 원리만을 이해하고 있다면 LightGBM의 소스를 실행하는 것만으로도 결과치는 얻을 수 있기 때문이다. 그러나 이 학습 프로세스를 순서대로 고찰하는 것이 필요한 이유는 모델에 대한 평가, 즉 모델이 얼마나 훌륭한지를 평가하는 것이 아니라 내 학습 데이터가 모델과 잘 매칭되어 원하는 퍼포먼스를 얻을 수 있는가에 대한 평가이다.

알고리즘을 선정하는 과정은 최적의 결과를 얻을 수 있는 도구를 선택하는 과정이라면, 알고리즘을 평가하는 과정은 데이터가 원하는 형태로 학습되어 적절한 결과를 더 높은 확률로 얻기 위한 반복 학습을 동반한다. 세세한 프로세스 과정을 통찰하고 하위노드생성과정을 이해하고 나면 어떤 파라미터들을 세밀하게 조정해야 더 나은 결과값을 얻을 수 있는지 알 수 있다.

각각의 지수 산출 레이어의 학습 방법은 의사결정트리 GBDT 알고리즘으로 구현될 수 있었으나 이것을 변형하여 구현될 수도 있기 때문에 LightGBM 모델을 선택하였다. 학습이 수행되기 전에 복수의 하이퍼 파라미터 값들이 설정될 수 있다.

일례로 손실함수의 종류는 로지스틱 손실함수를 사용하였는데 전체 기업 각각의 특성 중 연속적인 특성에 해당되는 특성값을 구간별로 나누어, 이진 특성으로 변화하는 bucketing의 값은 255, 점진적인 학습을 통하여 최소 경사값을 찾는 경사 하강법 적용을 위하여 필요한 학습률은 0.15, 트리 생성 조건으로서 학습 과정에서 생성될 수 있는 최대 의사결정트리의 수는 1000, 의사결정트리의 가지 수는 59, 학습 과정에서의 과적

합을 줄이기 위하여 전체 데이터에서 학습에 사용되는 데이터를 무작위로 추출하는 경우 적용되는 추출비율인 bagging은 0.8로 설정될 수 있다. 이러한 하이퍼 파라미터 값들은 같은 연구를 하는 다른 연구자들에 의해서 다르게 설정될 수도 있을 것이다.

이는 결과값에 대해서 평가를 하는 과정에서도 얼마든지 바뀔 수 있으며 특히 기계학습의 최대 난제인 과적합의 문제를 해결하기 위해서도 다양한 변수 세팅이 필요할 수 있다.

또 다른 예로서 상기 의사결정트리는 하위노드가 생성되는 각각의 단계에서, 하위노드가 없는 노드들 각각에 포함된 학습 데이터 집합 각각에 대하여 손실함수를 참조하여 손실값을 각각 계산하고 계산된 손실값이 가장 큼으로써 해당 학습 데이터의 집합을 분할하여 그에 대응되는 하위노드를 생성할 때 정보 이득이 가장 큰 특정 노드에 대해서만 하위노드를 생성하는 leaf-wise 트리 성장방식이 적용될 수 있다.

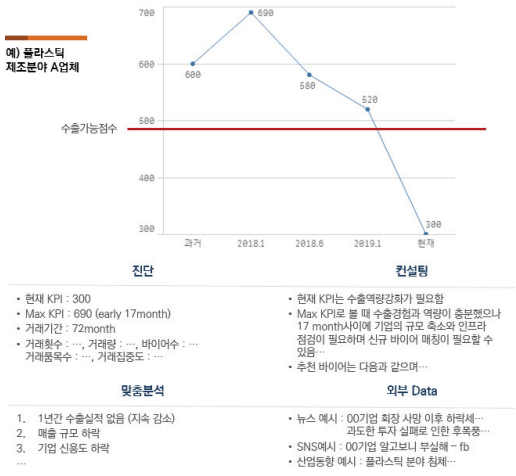
3.3.3 학습 데이터와 검증 데이터 비율

본 연구에서는 모델무역수출지수 산출 레이어, 신용등급지수 산출 레이어 및 현금흐름 등급 산출 레이어 각각의 학습 과정에서 제 1 트리 혹은 제 m 트리 각각의 생성 시마다 입력되는 학습 데이터로서의 제 1 학습 데이터, 제 2 학습 데이터 등 각각은, 전체 기업 각각의 기업수출 기초정보 전체 중 설정된 비율에 의하여 무작위로 추출된 부분집합의 데이터일 수 있다. 이때 설정된 비율은 하이퍼 파라미터로 정해진 bagging 비율이 적용되어 그에 따라 이루어지는 것일 수 있다.

예를 들어, 전체 기업이 1000개라고 가정할 때 bagging 값이 0.8로 설정되었다면 트리 각각의 생성 시에 root node에 대응되어 입력되는 학습 데이터는 800개의 데이터가 무작위로 선택될 수 있는데, 복원표본 샘플링으로 데이터를 선택함으로써 800개의 데이터 중 적어도 일부의 데이터는 특정 기업에 대한 데이터가 중복하여 선택될 수 있으므로, 800개의 데이터에 대응되는 기업의 개수는 800개보다 적을 수 있다.

Ⅳ. 학습 결과의 활용

학습 결과로서 나온 0-1사이의 값들은 실제 수출 여부를 대입하여 0.87 이상의 기업들을 수출가능성이 있는 기업으로 분류하였다. 여기에 무역수출 라이브 지수 개념을 적용하기 위하여 시계열 분석 방법을 적용하였다. 이는 <그림 4>를 참조한다.



〈그림 4〉 무역수출 라이브 지수 실 예

이는 실제로 서비스되고 있는 컨설팅의 서비스 컨셉을 도식화한 자료이다. 학습 데이터로 활용된 2017년 데이터를 과거 데이터로 간주하고 2018년을 전반기와 하반기로 나누었다. 다시 2019년을 상반기와 컨설팅이 이루어지고 있는 현재 시점으로 배치하였을 때 5개의 점으로 표시할 수 있다.

첫 번째 진단으로서 현재 KPI가 매우 낮은 것을 확인할 수 있다. 1차적인 이 분석만으로 분류

할 때 이 기업은 수출 불가 판정을 받을 수밖에 없을 것이다.

그러나 Max KPI는 현재로부터 17개월 전에 업계 최고치를 기록하였으며 거래 기간이 72개월이 넘는 상당히 우수한 기업이었음을 알 수 있다.

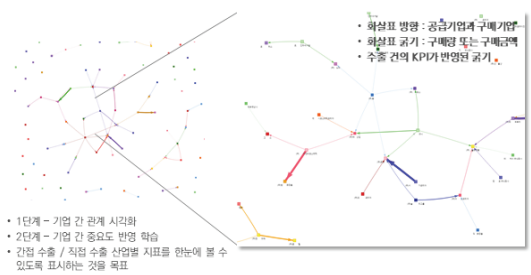
이에 대한 컨설팅으로서 현재 KPI는 수출역량 강화가 필요하고 17개월 동안 계속해서 하락하고 있는 것이 기업의 규모 축소나 인프라 점검에 대한 진단이 필요하다고 판단할 수 있다. 신규 바이어 매칭이 시급하며 추천 바이어에 대한 정보를 연계할 수도 있을 것이다.

이러한 맞춤 분석으로서 외부 데이터를 수집하여 융합할 때 뉴스, SNS, 산업 동향 등의 정보를 크롤링하여 기업의 위기와 대응해서 해석할 수 있는 여지도 존재한다.

다시 무역수출지수를 산출하고자 했던 취지로 돌아와서 수출이 가능한 기업을 발굴하고자 할 때 <그림 5>와 같은 시각화 분석이 가능하다.

1단계로 기업 간의 관계를 시각화하고 2단계로서 기업 간 중요도를 반영한 학습을 반영하며 간접 수출과 직접수출 산업별 지표를 한눈에 볼 수 있도록 표시하는 것을 목표로 한다. 이 차트에서 화살표 방향은 공급기업으로부터 구매기업을 표시하는 것이며 화살표의 굵기가 구매량 또는 구매금액의 규모를 나타낸다. 본 연구에서 산출된 KPI가 반영된 굵기라고 표현할 수 있다.

네트워크 차트를 통하여 표현하고자 하는 것은 어떤 기업을 중심으로 볼 수 있는 그 업계의 생태계이다. 자동차를 수출하고 있는 A 대기업에 납품하는 부품생산업체, 부품에 들어가는 나사를 생산하는 업체 등 계속해서 이어지는 공급 사슬 혹은 공급 네트워크망을 표현할 수 있는 자료로 활용될 수도 있을 것이다.



〈그림 5〉 수출 기업간 네트워크 차트

Ⅴ. 결론 및 한계점

궁극적으로 무역수출 라이브 지수는 구매확인서와 내국신용장의 데이터를 중심으로 분석하여 기업의 역량을 판단할 수 있는 다양한 지표들을 학습하여, 학습되지 않은 새로운 기업정보를 입력하였을 때 수출 가능성 여부를 판단하는 지표이다. 이를 토대로 수출 가능성이 있는 기업들이 어떤 역량을 보강하여 수출로 나아갈 수 있는지를 판단하는 지표로도 활용될 수 있다.

응용하여 직접 수출하지 않는 기업들보다 규모도 작고 신용등급도 낮지만 다른 어떤 요인에 의해서 잘 수출하고 있는지 요인을 분석하는 도구가 될 수도 있다.

반면에 이 연구는 특정 산업 분야의 특성을 반영하지 못하는 한계점을 가질 수 있다. 수출 규모가 작지만 다량의 생산품을 더 많은 바이어에게 판매하고 있는 기업은 본 연구의 학습에 의하면 매우 높은 점수를 받을 수 있지만 수출 규모만 큰 기업들과 같은 점수를 받을 수도 있다. 이에 다양한 산업 분야의 전문가들이 이 수출지수를 이해하고 사용하여 보정 할 수 있는 추가 연구가 필요할 것으로 사료된다.

참 고 문 헌

[1] Guolin Ke. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA. NIPS 2017.

[2] Lee, B. M. An Influence of the Fourth Industrial Revolution on International Trade and Countermeasure Strategies to Promote Export in Korea. *Korea Trade Research Association*. 2017.06. 1-24.

[3] Guolin Ke. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Data-science. 2018.06.

[4] Juanjuan Li. A hierarchical framework for ad inventory allocation in programmatic advertising markets. *Electronic Commerce Research and Applications* 31. 2018. 40-51.

[5] Yung-Chia Chang. Application of eXtreme gradient boosting trees in the construction of credit risk assessment models for financial institutions. *Applied Soft Computing Journal* 73. 2018. 914-920

[6] Yung-Chia Chang. Combining Principal Component Analysis, Discrete Wavelet Transform and XGBoost to trade in the financial markets. *Expert Systems With Applications* 125. 2019. 181-194.

[7] Wen Zhang. DeRec: A data-driven approach to accurate recommendation with deep learning and weighted loss function. Electronic Commerce *Research and Application*s 31. 2018. 12-23.

[8] Yang Zhao. How knowledge contributor characteristics and reputation affect userpayment decision in paid Q&A? An empirical analysis from the perspectiveof trust theory. *Electronic Commerce Research and Applications* 31. 2018. 1-11.

[9] Cheng Chen. LightGBM-PPI: Predicting protein-protein interactions through LightGBM with multi-information fusion. *Chemometrics and Intelligent Laboratory Systems* 191. 2019. 54-64.

[10] Jianing Niu, Yongmei Sun. Noise-suppressing channel allocation indynamic DWDM-QKD networks using LightGBM. *Optics Express Journal* 27. 2019.10. 22-28

[11] Kim, B. H. (2011), The Impact of FinTech on Transactions in International Trade. Korea Trade Research Association. 2018.04. 126-157.

[12] Baoguang Tian. Predicting protein-protein interactions by fusing various Chou's pseudo components and using wavelet denoising approach. *Journal of Theoretical Biology* 462. 2019. 329-346.

[13] Muhammad Tahir. Sequence based predictor for discrimination of enhancer and their types by applying general form of Chou's trinucleotide composition. *Computer Methods and Programs in Biomedicine* 146. 2017. 69-75.

[14] Xiaojun Ma. Study on a prediction of P2P network loan default based on the machine learning LightGBM and XGboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications* 31. 2018. 24-39.

[15] Wei Ou. Training attractive attribute classifiers based on opinion features extracted from review data. *Electronic Commerce Research and Applications* 32. 2018. 13-22.

[16] Xiaowen Cui. UbiSitePred: A novel method for improving the accuracy of ubiquitination sites prediction by using LASSO to select the optimal Chou's. pseudo components. *Chemometrics and Intelligent Laboratory Systems* 184. 2019. 28-43.

저 자 소 개

김 희 천(Heecheon Kim)

·2011년 : 명지대학교 컴퓨터공학과 (학사)
·2015년 : 연세대학교 공학경영 (석사)
·2019년 : 연세대학교 융합기술경영공학 (박사수료)
·현재 : ㈜미소정보기술 책임연구원
·관심분야 : 빅데이터 분석 및 컨설팅, 신사업 발굴 및 기획, 의료데이터 융합 분석, 인공지능 모델 개발

임 춘 성(Choon Seong Leem)

·1985년 : 서울대학교 산업공학과 (학사)
·1987년 : 서울대학교 산업공학과 (석사)
·1992년 : Univ. of California at Berkeley (박사)
·1993년~1995년 : 미국 Rutgers University 산업공학과 조교수
·현재 : 연세대학교 산업공학과 교수
·관심분야 : 비즈니스 모델(BM) 개발, 신기술 융합 서비스 모델 개발, 산업경쟁력 평가개발

성 주 원(Juwon Sung)

·2015년: 고려대학교 디스플레이반도체물리학과 (학사)
·현재 (주)미소정보기술 전임연구원
·관심분야: 빅데이터, 인공지능, 컴퓨터비전, NLP, 클라우드기술