가상화된 관세청 수입신고 데이터셋 소개

Introducing synthetic import declaration datasets

기초과학연구원 데이터사이언스그룹

May 20, 2022

Abstract

관세청의 수입 신고 데이터는 개인 정보, 무역 정보 등 데이터의 민감성 이슈 등으로 인하여 내부적으로만 활용되고 있다. 본 연구팀은 관세 행정 분야의 도메인 지식을 확산시키고, 우범 선별 및 관계성 예측 등 관세 분야의 연구 저변을 넓히기 위한 취지로, 가상화된 일반통관 수입신고서 데이터셋을 제작 및 배포한다. 적대적 생성 신경망을 이용해 만들어진 데이터셋은 54,000건으로 21개의 속성을 가진 수입 신고서 형태로 이루어져 있으며, 속성값의 분포, 관계성 측면에서 실 데이터셋과 상당히 유사하다. 또한 데이터를 구축하는 과정에서 원 데이터에 존재하는 속성 간의 의존성 및 경향성을 유지하기 위하여 활용한 방법을 소개한다.

Keywords— 수입신고 데이터, 가상 데이터 생성, 태뷸러 데이터, 우범 선별, 관계성 분석

1 서론

수입 통관이란 수입하고자 하는 물품을 세관에 신고하고, 관세법 및 기타 법령에 따라 신고가 적법하고 정당하게 이루어진 경우에 이를 수리하고 관세의 납부와 함께 물품이 반출될 수 있도록 하는 일련의 과정을 말한다. 통관 과정에서 수입자 (화주)나 관세사가 작성하는 수입 신고서의 정확도가 굉장히 중요하며, 신고서 내용에 오류가 존재할 경우 정정조치와 함께, 필요시 물품을 검사한다. 따라서, 관세청은 수입 신고 정확도를 향상하기 위하여 HS별「수입신고 품명 · 규격작성 가이드라인」을 준수하도록 당부하고 있다.

150달러 이상의 물건을 해외에서 구매할 경우 수입 신고의 대상이 되고, 수입자는 전자통관시스템(UNI-PASS, http://unipass.customs.go.kr)을 이용하여 직접 신고할 수 있다. 하지만, 정식 수입신고의 경우 전문적인 용어 및 신고항목이 많아 소비자가 직접 처리하기에는 부담스러운 부분이 많다. 따라서, 대부분의 사업자는 전문적인 업무지식을 가진 관세사를 통하여 업무를 대행하고 있으며, 일반인의 경우 20%의 고세율을 감내하고 간이 신고를 진행하기도 한다. 신고하는 물품의 품목에 따라 세율이 다르기 때문에 세관에서도 수입 신고 내용에 대한 철저한 분석을 거치고, 신고서 정보를 바탕으로 우범 가능성을 예측하여 검사 대상 물품을 선별한다 [1]. 수입 신고서 정보는 개인 정보 및 물품의 가격 정보가 포함되어 있는 이유로, 민감한 데이터로 분류되어 외부에 공개하지 않는다. 따라서, 이를 활용한 연구 및 분석은 유관 부서 이외에는 진행하기 어려웠다.

개인 직구등 자가 수입 통관을 원하는 사업자 및 일반인의 수입 통관 및 신고 절차에 대한 이해도를 높이는 동시에, 잘못된 신고로 비롯된 통관 효율성 저하를 방지하는 일환으로, 본 논문에서는 적대적 생성망 기법 CTGAN을 활용하여 제작한 수입 신고 가상 데이터를 소개한다.[1] 데이터셋은 54,000 건으로, 각 건은 수입 신고서 형태로 이루어져 있으며, 속성값의 분포, 관계성 측면에서 실 데이터셋와 상당히 유사하다. 이에 힘입어 본 데이터는 우범

[1] http://xxx.yyy

Table 1: 데이터 명세 - 링크 확인

속성	예시값	의미
신고일자	2020-01-01	신고서가 제출된 날짜
신고세관부호	13	세관의 부호
수입신고구분코드	B	일반 및 간이신고여부 등 수입신고의 종류 부호
수입거래구분코드	11	일반수입, 수탁가공무역 등 수입거래의 종류 부호
수입종류코드	21	내수용, 수출용원재료 등의 수입 종류 부호
징수형태코드	11	징수의 종류에 따른 부호
운송수단유형코드	10	운송수단 및 운송용기 부호
신고인부호	L77JJEG	신고인 상호와 성명에 따른 암호화된 부호
수입자	HQ0W7JA	수입자 통관고유부호
해외거래처부호	PBP2MYI	해외거래처 상호의 부호
특송업체부호	MWIDNS	특송 방법 및 업체의 일련번호
HS10단위부호	8407210000	10자리 물품 분류 코드
적출국가코드	JP	신고물품의 해외선적국가 부호
원산지국가코드	JP	원산지 국가의 부호
관세율	8.0	해당 품목의 세율 (%)
관세율구분코드	A	해당 품목의 세율 구분 부호
원산지표시유무코드	B	원산지 표시유무 및 표시면제사유 등
신고중량	1262.0	물품의 포장용기를 제외한 중량 (KG)
과세가격원화금액	1437418.0	구매자가 실제로 지급한 금액 (원)
우범여부	1	우범 여부 (0/1)
핵심적발	1	가중치가 높은 우범 여부 (0/1)

선별, 품목 분류, 물동량 예측 등 관세 분야의 연구에 활용될 수 있다. 관세청은 벤치마크 데이터를 활용하여 3개의 학교 (숭실대, 충남대, 카이스트)를 대상으로 우범 선별 경진대회를 진행하는 동시에, 대규모의 실 데이터에 접근할 수 있는 안심 구역을 청 내에 마련하여 관세행정업무에 대한 도메인 지식을 확산시키고, 실제 업무에 적용 가능한 수준의 AI 알고리즘을 발굴하기 위한 데이터 개방을 추진하고 있다.

2 데이터 소개

2.1 데이터 스키마

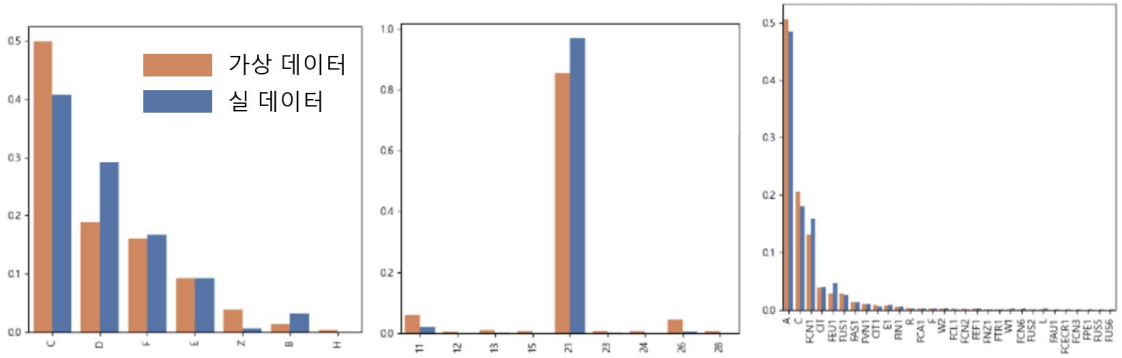
본 데이터는 총 54,000건의 수입 신고 건으로 이루어져 있으며, 각 건은 물품 단위의 수입 신고 내용을 모사한다. 수입 신고서[2]에 기입되는 62가지의 속성 중 업무적으로 가장 중요한 대표값 21개가 데이터를 구성한다. 중복된 의미를 갖는 속성들은 가상 데이터를 만드는 과정에서 제외하였고 독립 변수를 중심으로 대표값을 선정하였다. 표 1은 데이터 명세 및 예시값을 보여준다. 관세청에서 제공하는 무역통계부호[3]를 활용하여 데이터에 존재하는 범주형 변수 (Categorical variable)의 고유값들에 대한 정보를 알아볼 수 있다. 업무 프로세스를 고려하면, 21개의 속성은 신고 단계에서 작성된 내용과 검사/반출 단계에서 작성된 내용으로 구분할 수 있다. 검사/반출 단계에서 작성된 속성은 우범여부 및 핵심적발이 있는데, 우범여부 속성은 물품 검사 결과 신고된 내용과 실제 물품이 상이한 우범으로 판단되었을 때 기입한다. 우범 중에서도 세관에서 중점 관리하는 위반항목인 상표권/저작권 침해물품, 국민건강 위해물품, 원산지위반 등에 해당되면 핵심적발로 표시한다.

[2]수입 신고서 서식 및 작성 방법: `https://bit.ly/import-declaration-form`

[3]무역통계부호: 수출입 신고서를 작성할 때 필요한 통계부호로서 우리나라 이해관계자가 수출입물품에 대한 신고를 위해 신고서를 작성할 때 필요하다. `https://www.data.go.kr/data/3040477/fileData.do`

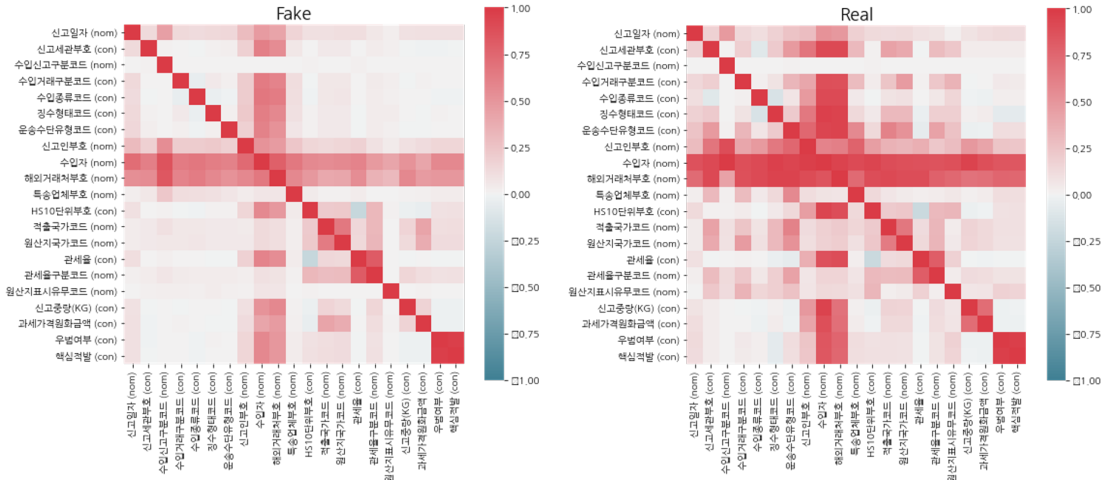
2.2 데이터 분석

생성한 가상 데이터의 신빙성 확인을 위하여, 실 데이터와의 속성값 분포 비교를 진행했다. 그림 1은 데이터 대표적인 속성값—수입통관계획코드, 적출국가코드, 통관지세관부호의 분포가 두 데이터셋 간 유사하다는 점을 보인다. 더불어 데이터의 유사성을 확인하기 위한 통계적 검증을 수행하였다. “신고인부호”, “HS10단위부호” 등 불연속적인 값을 가진 속성들은 Chi-Square test로, “관세율” 및 “과세가격원화금액” 등의 연속적인 값을 가진 속성들은 Two-sample Kolmogorov-Smirnov test로 유사성을 검증하였다. 유사도는 -1에서 1 사이의 값으로 정의되며, 두 컬럼이 같은 분포로부터 샘플링되었을 때 유사도 값이 1에 가까워진다. 종합해 보았을 때, 가상 데이터 내 컬럼 간의 경향성이 원본 데이터의 경향성에 비해 조금 약하지만, 전반적으로 유사하다는 점을 그림 2를 통해 알 수 있다. 그림 3은 다운스트림 테스크인 우범 선별을 진행하는 과정에서 얻어진 속성별 feature importance를 통해 두 데이터의 유사도를 판단한다. 두 데이터 모두 수입자부호, 과세가격원화금액, 신고중량, 신고인부호, HS10단위부호 등이 중요한 속성값으로 나타난다.



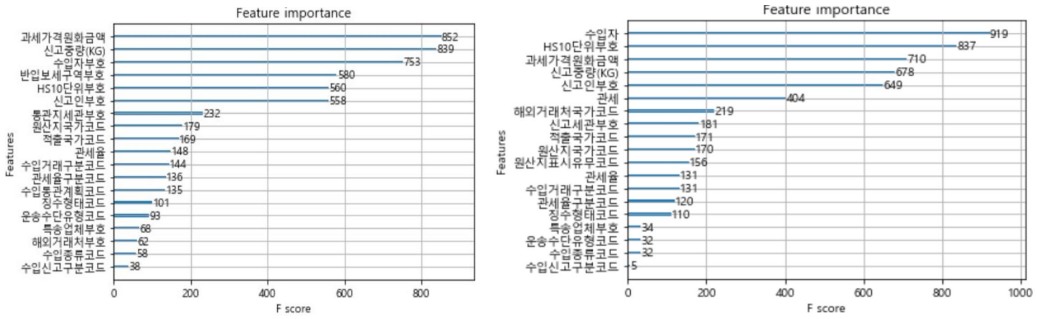
(a) 수입통관계획코드 (b) 적출국가코드 (c) 통관지세관부호

Figure 1: 대표 속성들로 알아본 가상 데이터와 실 데이터의 유사도 비교



(a) 가상 데이터의 속성 간 관계성 (b) 실 데이터의 속성 간 관계성

Figure 2: 속성 간 관계성을 통한 가상 데이터와 실 데이터의 유사도 확인



(a) 가상 데이터 (b) 실 데이터

Figure 3: 우범 선별 문제에서의 feature importance를 통한 두 데이터셋 비교

3 데이터 생성 과정

3.1 가상 데이터 생성에 앞선 실 데이터 전처리 및 익명화

2020년 1월 1일부터 2021년 6월 30일 사이 18개월간 신고된 일반 통관 수입신고건 2,469만 건 중 검사선별이 진행된 41만 건의 데이터 중, 무작위로 54,000개의 행을 추출하여 가상 데이터 제작에 활용하였다. 가상화 작업에 투입하기 전, 데이터의 모든 개인 정보는 제외되거나 익명화되었으며, 민감 정보중 하나인 건별 가격 정보 역시 가우시안 노이즈가 추가된 품목별 단가로 대체하였다. 수입신고 데이터의 특성상 속성 간 관계성 유지가 가능한 범위 내에서 익명화 및 가상화를 진행하였으며, 무역통계부호를 참고하여 생성된 데이터의 해석이 가능하다.

3.2 CTGAN을 이용한 가상 데이터 생성

데이터의 가상화를 위해 태뷸러 데이터에 특화된 적대적 신경망 알고리즘을 활용하였다. The Synthetic Data Vault (SDV) 라이브러리에 구현된 적대적 신경망 알고리즘 TGAN [2], TVAE [3], CTGAN [3] 중 가장 넓은 생성 커버리지를 보인 CTGAN [3]을 채택하였다. 리소스의 한계를 극복하기 위해 월별 데이터를 샘플링한 후 CTGAN 알고리즘을 통해 가상 데이터를 생성하여 병합하는 형태로 전체 기간의 가상 데이터를 생성하였다.

3.3 컬럼 간의 관련성 유지를 위한 테크닉

태뷸러 데이터에는 속성 간의 관련성이 존재하며, 특히 수입 신고 데이터의 경우, 납득 가능한 형태가 되려면 관계성이 철저하게 유지되어야 한다. CTGAN이 태뷸러 데이터에 특화된 적대적 신경망임에도 불구하고, 원본 데이터를 그대로 이용할 경우 21개의 속성 간의 명확한 종속 관계가 깨지는 경우가 발생하였다. 예를 들면, “HS10단위부호”—“적출국가코드”—“원산지국가코드”—“관세율”—“관세율구분코드” 속성은 서로 연동되어 있으며, 생성된 데이터에서도 이 관계성이 항상 유지되어야 한다. 또한 “핵심우범” 컬럼은 우범 중에서도 가중치가 높은 우범인지의 여부를 의미하므로 “핵심우범” 컬럼의 값이 1(참)인 경우 “우범여부” 컬럼의 값은 반드시 1이어야 한다. 이렇듯 강한 연관성을 가진 컬럼들 간 관계를 유지한다는 조건을 만족하기 위해, 연관 관계에 있는 변수를 결합하여 하나의 긴 값을 가지는 범주형 변수를 만들었다. 이후 CTGAN을 수행하여 만들어진 가상의 값을 다시 분리하는 방식으로 생성된 데이터의 속성간 관계를 유지하였다. “과세가격원화금액”와 같은 속성도 중량 및 물품의 종류에 따라 결정되는 수치형 변수로, HS10코드로 표시되는 물품의 종류별 단가의 평균값에 총 중량을 곱하여 재생성하였다.

4 활용 방안—가상 데이터를 활용한 우범 선별

수입 신고 데이터에 존재하는 우범 여부/핵심 적발을 타겟 변수로 하여 통관 시스템의 효율화를 위한 우범 선별 문제를 접근할 수 있다 [1, 4]. 본 섹션에서는 문제의 소개와 함께, 가상 데이터과 원본 데이터 간 우범 예측 성능이 얼마나 유사한 경향을 보이는지 살펴보고자 한다.

4.1 우범 선별의 목적 및 문제 정의

마약이나 무기와 같은 위해 물품이나 거짓 정보로 신고된 부정 거래는 사회에 혼란을 야기하고 국민 안전에 치명적인 위협이 될 수 있다. 이러한 위험을 차단하기 위해, 관세 당국에서는 신고화물을 분석하고 우범 화물이 높은 물품을 정확히 예측하고 선별할 수 있는 기술의 개발을 필요로 하고 있다. 한정된 검사 인력을 가장 효율적으로 활용하기 위해 세관에서는 우범의 가능성이 높은 물품을 위주로 검사를 수행하고 있다. 이러한 현실적 한계를 반영하기 위해, 우범 선별 문제는 단순히 이진 분류의 정확도를 측정하기보다 우범 가능성이 높은 수입 신고의 집단인 "의심 집단"을 정의한 후, 이 중에서 실제 우범 화물의 비율을 성능 지표로 본다. 마찬가지로 수입 신고 데이터에서 "우범여부" 열을 타겟 변수로 설정하여 문제를 접근할 수 있다. 우범 여부와 관련 없는 "신고번호", "신고일자" 및 주요 우범 여부를 나타내는 "핵심우범" 열을 제외한 나머지 속성들을 우범 선별 모델의 학습에 활용하였다.

4.2 셋업 및 평가 방법

기존의 데이터를 바탕으로 새롭게 신고된 물건의 우범 여부를 예측하는 문제의 특성 상, 신고 일자를 기준으로 데이터를 나누었다. 2020년 1월부터 2020년 12월까지 신고된 데이터로 모델을 학습하였으며, 2021년 1월부터 2021년 3월까지의 데이터는 모델을 검증하는 데 사용하였고, 2021년 4월부터 2021년 6월까지의 데이터는 성능 평가를 위해 활용되었다. Logistic Regression, Decision Tree, Random Forest, AdaBoost, LightGBM [5], CatBoost [6], XGBoost [7] 등 다양한 모델을 활용하기에 앞서 범주형 변수는 label-encoding, 수치형 변수는 min-max scaling을 통한 전처리를 수행하였다. 모델은 각 물품에 대한 우범 가능성을 0과 1사이의 값으로 도출하며, 전체 물품 중 우범 가능성이 가장 높다고 판단되는 물품 n%을 검사하게 된다. 평가 지표로는 5%, 10%의 물품을 검사하였을 때의 적발율 (Precision; 검사 대상으로 선별된 물품 중 실제 우범 으로 판단된 비율)을 활용하였다.

4.3 실 데이터, 가상 데이터 간 모델별 성능 비교

위와 같은 방법으로 우리가 생성한 가상 데이터와, CTGAN에서 원본 데이터로 사용하였던 실제 데이터에서의 성능을 비교하였다. 표 2는 가상 데이터와 실 데이터에서 검사율에 따른 우범 선별 모델의 성능을 보여준다. 검사율이 낮을수록 우범 가능성이 높은 물품 위주로 검사가 이루어지기 때문에, 두 데이터셋 모두 검사율이 5%인 셋업에서의 적발율이 10%에서의 적발율에 비해 높다는 점을 확인할 수 있다. 또한 트리 기반 그래디언트 부스팅 모델인 CatBoost, XGBoost, LightGBM의 성능이 기초적인 모델들에 비해 높다는 점을 확인할 수 있었으며, 실 데이터에 비해 가상 데이터에서 모델 간 성능 차이가 더 두드러지게 나타나기도 하였다. 이로써, 가상 데이터를 테스트베드로 삼아 고도화된 우범 선별 모델의 개발이 가능하다는 점을 확인해볼 수 있었다.

5 디스커션

제작한 데이터셋에 관하여 알아둘 만한 부분을 정리하면 다음과 같다.

- 가상화를 진행한 이유: 신고자부호, 물품 분류 코드, 적출국가코드 등의 익명화를 진행하더라도, 전체 분포 속에 숨어 있을 수 있는 패턴이 간접적으로 노출될 수 있는 우려가 있어 CTGAN을 활용한 가상화를 진행한 데이터를 제작하게 되었다.

Table 2: 알고리즘별 우범률

	가상 데이터		실 데이터	
우범 선별 모델 — 검사율	n = 5%	n = 10%	n = 5%	n = 10%
Logistic Regression	0.2870	0.2657	0.2320	0.2237
AdaBoost	0.3722	0.3363	0.5280	0.5047
Decision Tree	0.3722	0.3520	0.5440	0.4993
Random Forest	0.4215	0.3980	0.5520	0.5220
CatBoost	0.5874	0.5000	0.5813	0.5206
XGBoost	0.5942	0.5067	0.5760	0.5113
LightGBM	0.7220	0.5897	0.5493	0.5220

- 데이터의 분포: 사용자의 편의를 위해, 통관 절차 중 세관원을 통해 검사가 진행된 물건을 대상으로 가상 데이터를 생성하였다. 따라서 본 데이터의 경우 모든 물품의 우범여부 / 핵심적발 라벨이 존재한다. 국가별 검사 기준에 따라, 실제로는 검사를 거치지 않고 통관 절차를 거치는 물품이 존재하며, 해당 물품은 우범여부 / 핵심적발 라벨이 존재하지 않는다. 일부 데이터의 라벨을 지우는 등의 후처리를 통하여 이런 상황을 시뮬레이션해 볼 수 있다.
- 더 나은 가상화 방안: 최근 이미지 생성 분야에서 GAN을 뛰어넘는 모델로 diffusion model 등이 거론되고 있다 [8]. 태뷸러 데이터에 특화된 diffusion model이 개발되면, 데이터 가상화에 활용해 볼 수 있을 것이다.
- 데이터 활용 방안: 본 수입 신고 데이터는 우범 선별 문제 이외에도 HS6단위 부호를 예측하는 품목 분류 문제 [9]나 코드별 무역 패턴 분석, 수입자, 신고인, 거래처 간의 상관관계분석 등 다양한 관세 행정 문제에 활용될 수 있을 것이다. 또한 데이터에서 학습된 패턴을 활용하면 개인의 수입 신고 프로세스를 편하게 해주는 서비스 등이 가능할 것이라 기대한다.

6 결론

본 연구에서는 관세 행정 분야의 도메인 지식을 확산시키고, 연구 저변을 넓히기 위한 일환으로 제작한 가상화된 일반통관 수입신고서 데이터를 소개하였다. 태뷸러 데이터에 특화된 적대적 신경망 알고리즘 CTGAN을 통해 데이터를 제작하였으며, 무역통계부호 등의 카탈로그를 활용하여 데이터의 해석이 가능하다. 공개하는 데이터는 대표적인 속성값의 분포가 실 데이터와 유사하며, 속성 간 관계성이 마찬가지로 존재한다. 본 데이터를 활용하여 기계 학습 기반 우범 선별 알고리즘, 품목 분류 알고리즘 등 통관 절차의 효율화에 필요한 선행 모델을 개발 및 고도화할 수 있으리라 기대한다.

References

[1] S. Kim *et al.*, "DATE: Dual Attentive Tree-Aware Embedding for Customs Fraud Detection," in *KDD*, 2020.

[2] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," *arXiv preprint arXiv:1811.11264*, 2018.

[3] L. Xu, M. Skoularidou, A. Cuesta-Infante, and K. Veeramachaneni, "Modeling Tabular data using Conditional GAN," in *NeurIPS*, 2019.

[4] S. Kim, S.-K. Song, M. Cho, and S.-H. Shin, "Transaction Pattern Discrimination of Malicious Supply Chain using Tariff-Structured Big Data," *The Journal of the Korea Contents Association*, 2021.

[5] G. Ke *et al.*, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree," in *NeurIPS*, 2017.

[6] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "CatBoost: Unbiased boosting with categorical features support," in *Advances in Neural Information Processing Systems 31*, Curran Associates, Inc., 2018, pp. 6639–6649.

[7] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *KDD*, 2016, pp. 785–794.

[8] P. Dhariwal and A. Nichol, "Diffusion models beat GANs on image synthesis," *arXiv preprint arXiv:arXiv:2105.05233*, 2021.

[9] E. Lee *et al.*, "Classification of goods using text descriptions with sentences retrieval," in *Korea Artificial Intelligence Conference (KAIA)*, 2021.