빅데이터를 활용한 국내 도서의 해외 판매시 굿셀러 예측

김나연, 김도영, 김미려, 정지영, 김현희
동덕여자대학교 정보통계학과
20181040@dongduk.ac.kr, 20181041@dongduk.ac.kr, 20181043@dongduk.ac.kr,
20191073@dongduk.ac.kr, heekim@dongduk.ac.kr

Prediction of Good Seller in Overseas sales of Domestic Books Using Big Data

Nayeon Kim, Doyoung Kim, Miryeo Kim, Jiyeong Jung, Hyon Hee Kim
Department of Statistics and Information Science,
Dongduk Women's University

요 약

한국 문학이 세계로 뻗어나감에 따라 해외 시장에서 자리를 잡는 것이 중요해진 시점이다. 본 연구에서는 2016 년도부터 2020 년도까지 최근 5 년간 해외 출간된 도서들 중에서 굿셀러로 분류되는 누적 5 천부 이상 판매 여부를 예측하고자 했다. 굿셀러로 분류되는 도서는 전체 번역 도서 중 적은 비율을 차지하여 데이터 불균형이 발생하였으며, 본 연구에서는 SMOTE 기법과 앙상블 알고리즘을 적용하여 데이터 불균형 문제를 해결하였다. 그 결과, 데이터 클래스 비율이 1:1 에 가까울수록 성능 개선 효과가 나타났으며 LightGBM 모델이 99.83%의 AUC 값을 얻어 다른 앙상블 알고리즘에 비해 가장 좋은 예측 성능을 보임을 검증하였다. 또한 누적 5 천부 이상 판매 여부 예측에 있어 큰 영향을 미치는 변수로는 작가가 가장 중요한 요인으로 나타났으며 출간 국가, 그리고 평점 평균, 평점 참여자 수 같은 온라인 요인도 판매 예측에 유의미한 변수로 나타난 것을 확인할 수 있었다.

1. 서론

해외에서 한국 문화의 인기가 높아지는 가운데 한국 문학도 다양한 문화권에서 주목받고 있다. 한국문학번역원에 따르면 한국 문학은 연평균 10%의 증가세를 보이고 있고 최근 번역원의 전체 지원건수 가운데 해외출판사가 한국문학 번역/출판을 일괄 신청하는 비중이 80%에 달한다. 이는 해외에서 한국문학을 출간하려는 자생적 수요가 증가했음을 나타낸다. 달리 말하면 이제 한국문학이 "문학 한류"의 초입에 서 있다고도 볼 수 있다. [1] 현재 해외시장에서의 늘어가는 수요에 부응하는 전략이 필요한 상황인만큼, 앞으로 해외 출간될 도서에 대한 판매량의 예측은 매우 중요하다. 따라서 본 연구는 한국문학번역원에서 진행한 해외 출간도서 판매현황 조사를 통해 누적된 데이터를 토대로, 리뷰, 작가, 책 정보를 수집해 한국문학 도서의 해외 판매 부수를 예측하기 위한 모델을 제시하고자 한다.

본 연구에서는 한국문학번역원에서 제공받은 최근 5 년(2016~2020 년) 누적 5 천 부가 판매된 해외 출간도서 데이터를 종속변수로 활용해 누적 5 천부 판매 여부 예측 모델을 제작했다. 5 천부를 기준으로 삼은 이유는 한국 문학 번역원의 "국내도서 초판 부수가 2 천~3 천 부 내외인 점을 고려, 판매부수 5 천 부는 해당 도서가 평균적인 초판 부수 이상 판매되었음을 보여주고 현지의 꾸준한 수요를 확인할 수 있는 지표로 볼 수 있다." 는 의견에 기반한다.

해당 데이터의 클래스가 약 1: 24.1 비율의 불균형 데이터인 관계로 Synthetic Minority Over-Sampling Technique(SMOTE) [2] 기법과 앙상블 알고리즘 [3]을 적용해 데이터 불균형 문제를 개선하고자 했다. 앙상블 알고리즘으로 Xgboost [4], GradientBoosting [5], Adaboost [6], Lightgbm [7], randomforest [8]를 적용해서 예측해본 결과, SMOTE 가 데이터 불균형 문제를 크게 개선하고 Lightgbm 이 다른 알고리즘과 비교하여 예

측 성능이 우수함을 보였다.

최종 선정 모델은 SMOTE 로 클래스의 비율을 1:1 로 조정해 LightGBM 으로 학습시킨 모델로, AUROC 값이 99.83%로 나타났다. 또한 어떤 요인들이 예측에 영향을 미쳤는지 분석한 결과, 작가의 해외 출간 횟수와 상위 출판 시장 규모를 가진 국가로의 출간, 평점평균, 평점 참여자수가 중요한 특징으로 확인되었다. 본 연구를 통해 앞으로 해외 출간될 국내 도서의 해외 경쟁력을 예측해볼 수 있고 더 나아가 보완하거나 필요한 요소를 인지하는 과정에 활용 가능하여 한국 문학의 해외 시장 성장을 넓힐 수 있는 발판이 될 것으로 기대된다.

논문의 구성은 다음과 같다. 2 장에서는 데이터 수집 및 전처리에 대해 다룬다. 3 장에서는 예측 모형을 제시하고 4 장에서 모형의 예측 성능을 비교 평가한다. 마지막 5 장에서는 본 연구의 결론과 시사점을 제시한다.

2. 데이터 수집 및 전처리

본 연구에서는 아마존(https://www.amazon.com/)과 굿리즈(https://www.goodreads.com/) 사이트에 등록이 되어있는 2016~2020 년 한국문학번역원의 지원으로 해외 출간된 도서 총 578 종을 대상으로 하였다. 도서 정보와 작가 정보는 DLKL(Digital Library of Korean Literature, https://library.ltikorea.or.kr/)에서 크롤링하여 수집하였으며, 아마존과 굿리즈에서 평점 정보와 2022 년 4 월 16 일까지 등록된 리뷰 총 26105 개를 크롤링하였다. 수상내역은 한국 문학 번역원으로부터 한국 문학 국제 수상 성과 자료를 전달받았다.

아마존과 굿리즈에서 크롤링한 데이터를 바탕으로 책 별로 평점 평균과 분산, 평점 참여자 수, 리뷰 수, 리뷰의 추천 수 등을 계산했다. 긴 리뷰 비율은 각 리뷰의 글자 수를 계산하여 전체 분포를 확인, 3 분위 수 이상을 긴 리뷰의 기준으로 정하여 작품별로 긴 리뷰의 비율을 구했다. 번역점수는 전체 리뷰 중 번역 관련 평이 포함된 문장들만 추출해 Vader nltk 로 감성분석을 진행한 뒤, 작품별로 compound 점수의 평균으로 번역 점수를 부여했다. 번역 점수가 부여되지 않은 결측치의 경우 번역에 관한 호평/혹평 모두 존재하지 않았다는 뜻이므로 평균값을 부여했다.

도서의 긍정/부정 리뷰의 비율은 문맥 고려가 가능한 언어 처리 모델인 Bidirectional Encoder Representations from Transformers(BERT) [9]를 이용했다. 1 점 리뷰를 부정, 5 점 리뷰를 긍정으로 분류한 학습 데이터로 파인 튜닝(batch_size=8, max_length=400)을 진행했으며, 그 결과 0.9644 의 예측 정확도를 보였다. 파인튜닝된 모델로 나머지 데이터에도 긍정/부정 라벨링을 진행해 각 도서의 긍정 및 부정 리뷰 비율을 구하였다.

DLKL 에서 크롤링한 데이터를 활용해 작품 별 번역 출간 수를 구했으며, 작가 해외 출간 횟수는 작가가 여러 명인 책의 경우, 가장 많이 해외에 출간한 작가를 기준으로 입력했다. 출판 시장 규모 상위 국가 출간 점수는 세계 지적 재산권 기구(WIPO)의 <World Intellectual Property Indicators 2021> 보고서를 참고하여 2020 년 연 수익을 기준으로 세계 출판 시장에서의 각 나라의 점유율을 계산하고, 상위 10 개국을 구했다. 그리고 상위 10 개국에 출간되었을 때마다 점유율을 점수를 주어 합산해 출판 시장 규모 상위 국가 출간 점수를 구했다.

3. 모델 생성

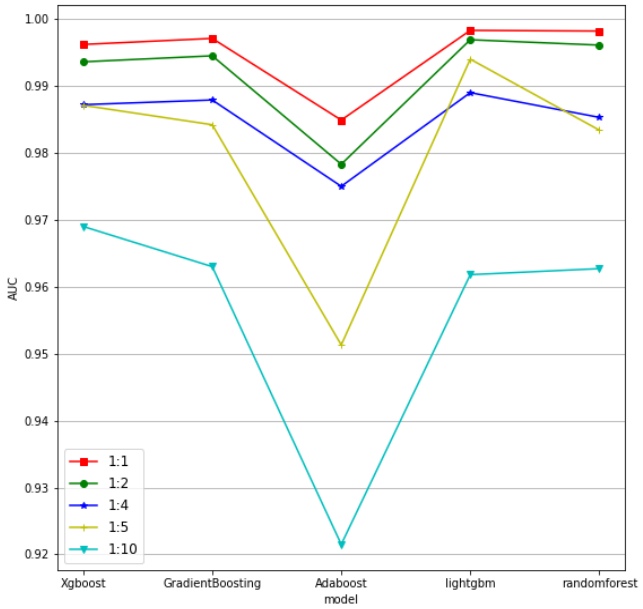
본 연구에서 사용된 누적 5 천부 이상 판매된 해외 출간 도서 데이터의 개수는 23 개이고, 그렇지 않은 데이터의 개수는 555 개로, 약 24 배 차이가 나는 불균형 데이터인 것을 알 수 있다. 불균형 데이터에 적용된 경우는 종속 변수의 균형을 맞추도록 데이터의 전처리 과정과 함께 앙상블 기법을 사용한 경우에서 분류모형이 효과적으로 적합된다는 사실이 알려져 있다. [10][11][12]

본 연구에서는, 데이터 불균형 문제를 해결하기 위해 SMOTE 기법을 적용한 데이터에 앙상블 알고리즘을 적용해 모델을 제작했다. 데이터 균형 비율에 따른 분류 결과를 비교하기 위해 각 범주의 데이터를 1:10, 1:5, 1:4, 1:2, 1:1 로 각각 구성해 진행하였으며, 앙상블 기법으로는 대표적인 부스팅 알고리즘인 Xgboost, GradientBoosting, Adaboost, Lightgbm 과 배깅 방법을 사용한 결정 트리의 앙상블 알고리즘인 randomforest 을 적용하여 예측 성능이 가장 뛰어난 최적의 모델을 찾고자 했다. 성능 평가는 5 겹 교차검증을 진행하여 다수 범주와 소수 범주 모두의 정확도를 고려할 수 있는 AUROC 를 사용하였다.

4. 실험결과

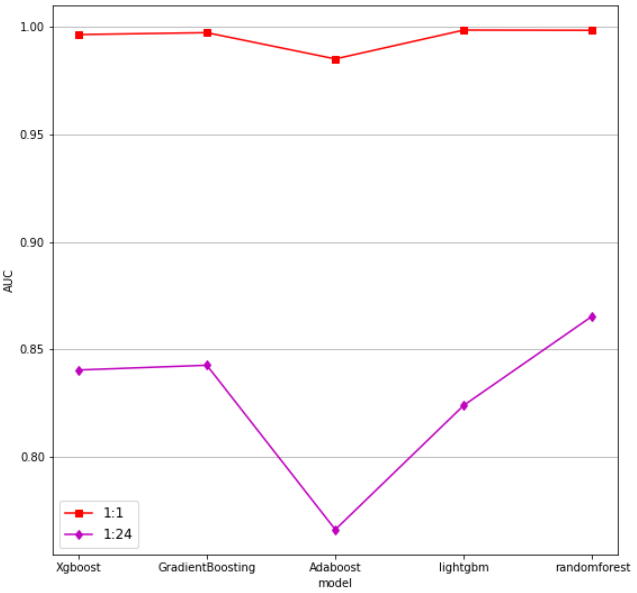
그림 1 은 클래스 구성 비율과 적용한 모델 별로 5 겹 교차 검증으로 평가한 성능을 비교한 그래프이다. 성능을 비교한 결과, SMOTE 기법을 적용해 각 범주 데이터의 비율을 1:1 에 가깝게 만들수록 모든 모델에서 전반적으로 예측 성과가 눈에 띄게 높아지는 것을 확인할 수 있었다. 또한, 모든 클래스 비율에서 Adaboost 모델의 성능이 가장 떨어졌으며, 클래스 비

율이 1:5 비율로 구성했을 때부터 Lightgbm 의 성능이 타 알고리즘과 비교해 성능이 가장 높게 나타났다.



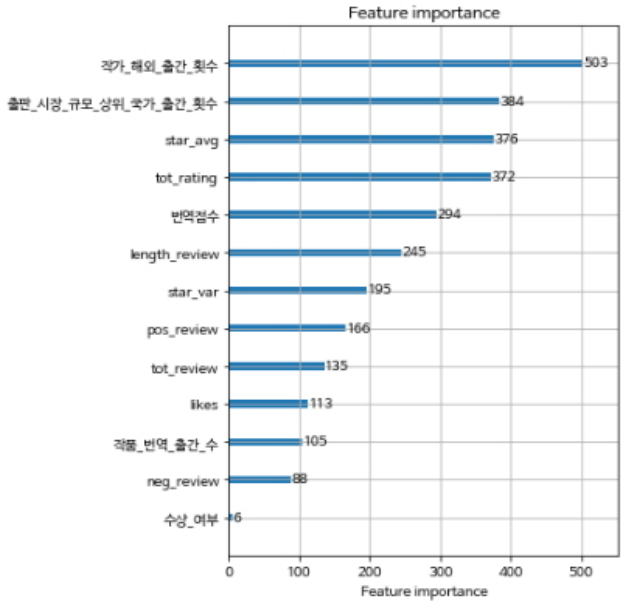
(그림 1) 클래스 구성 비율별 성능 평가

그림 2 는 SMOTE 를 적용 안 한 모델의 성능과 데이터의 비율을 1:1 로 조정한 모델의 성능을 비교한 그래프이다. SMOTE 기법을 사용해 데이터 불균형 문제를 해결한 것이 예측 성능을 크게 높였음을 확인할 수 있었다. 최종적으로는 클래스의 비율을 1:1 로 조정하여 Lightgbm 을 통해 예측한 결과가 0.9983 의 AUROC 값으로 가장 좋은 성능을 나타냄을 알 수 있었다.



(그림 2) SMOTE 적용 유무에 따른 성능평가

그림 3 은 최종 선택된 lightgbm 모형의 변수 중요도 결과이다. 그래프를 보면 해외에서의 누적 5 천부 이상 판매 여부를 예측하는 데 가장 중요한 변수는 작가의 해외 출간 횟수이다. 그 다음은 상위 출판시장 규모를 가진 국가로의 출간, 평점평균, 평점 참여자수가 중요한 특징으로 확인되었다. 번역에 대한 만족도나 긴 리뷰의 비율 역시 예측에 있어서 적지 않은 영향을 미친 중요 변수로 나타났다. 그리고 반대로 작품의 수상 여부는 영향이 미미한 것으로 확인되었다.



(그림 3) 최종 모델의 변수 중요도

5. 결론

본 논문에서는 웹 크롤링을 기반으로 수집한 리뷰 및 책, 작가 관련 정보 데이터를 기반으로 한국 문학 번역본의 해외 경쟁력을 예측 모델을 제안하였다. SMOTE 를 이용해 여러 샘플링 비율을 적용하고 여러 가지 앙상블 모델을 사용하여 성능을 확인한 결과, 소수 클래스의 데이터를 확장시키는 오버샘플링 방법과 함께 앙상블 기법을 사용하는 것이 그렇지 않은 경우보다 더 높은 성능을 얻었으며, 데이터 불균형 문제도 해결할 수 있음을 확인할 수 있었다. 그리고 최종적으로 선택된 모델은 클래스의 비율을 1:1 로 조정해 LightGBM 으로 학습시킨 모델이며, 성능 결과는 99.83%의 AUROC 로 우수한 성과를 보였음을 확인했다.

또한 중요도 그래프를 통해, 작가의 인지도가 해외 판매에 있어 가장 큰 요인이라는 것을 알 수 있었다.

출간된 국가의 출판시장 규모와 평점과 같은 온라인 요소도 도서의 해외판매를 예측하는 데 영향력이 있음을 확인할 수 있었다. 그리고 번역의 질 역시 중요한 요소로 밝혀졌다. 해외 출간에 있어서 저자의 의도를 살리고 해당 국가 독자들에게도 잘 전달되도록 번역하는 일이 실제 판매 예측에도 영향력이 있음을 확인할 수 있었다.

제안한 모델은 국내 도서의 해외 출간에 있어서 필요한 요소를 인지하는 과정에 활용하여 해외에서의 한국 문학의 성장에 도움을 줄 수 있을 것으로 기대한다. 앞으로 지속적인 해외 출간 판매현황 조사와, 더 다양한 데이터를 활용해 심층적인 분석을 한다면 해외 시장에서의 한국 문학의 경쟁력에 있어 큰 기여를 할 수 있을 것으로 기대된다.

본 연구에서는 불균형 데이터의 클래스의 비율을 맞추는 전처리 과정과 함께 적용했을 때 큰 효과를 볼 수 있는 것으로 알려져 있는 앙상블 알고리즘만을 적용하여 연구를 진행했다. 향후 연구에서는 앙상블 알고리즘뿐만 아니라 신경망 계열의 모델을 비롯한 여러 알고리즘을 적용해 실험을 진행할 것이다. 또한 변수 조합과 하이퍼파라미터 조절을 달리해서 성능을 비교해볼 예정이다.

참고문헌

[1] 뉴스페이퍼 (2022. 01. 18),「해외에서 가장 많이 팔린 한국문학은? 한국문학번역원 최근 5 년 해외출간 한국문학 판매현황 조사」, http://www.news-paper.co.kr/news/articleView.html?idxno=76610

[2] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall,W. Philip Kegelmeyer, "SMOTE: Synthetic Minority Over-samplingTechnique", Journal of Artificial Intelligence Research 16 (2002) 321-357

[3] Aurelien Geron,「핸즈온 머신러닝(2 판)」, 박해선 역 (2020), 한빛미디어.

[4] Tianqi Chen, Carlos Guestrin, "XGBoost: A Scalable Tree Boosting System", ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp. 785-794, Aug. 2016.

[5] Leo Breiman, "ARCING THE EDGE", Technical Report 486, Statistics Department, University of California, Berkeley, June. 1997.

[6] Yoav Freund, Robert E. Schapire, "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting", journal of computer and system sciences 55, 119-139, 1997

[7] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, Tie-Yan Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", 31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

[8] Leo Breiman, "Random Forests", Machine Learning, 45, 5–32, 2001

[9] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.

[10] Chawla, N. V., Lazarevic, A., Hall, L. O. and Bowyer, W. K. (2003). Smoteboost: Improving prediction of the minority class in boosting, Seventh European Conference on Principles and Practice of Knowledge Discovery in Databases, 107-119.

[11] Seiffert, C., Khoshgoftaar, T., Van Hulse, J. and Napolitano, A. (2010). Rusboost: A hybrid approach to alleviating class imbalance, Institute of Electrical and Electronics Engineers, 40, 185-197.

[12] Liu, X., Wu, J. and Zhou, Z. (2009). Machine learning for the detection of oil spills in satellite radar images, Institute of Electrical and Electronics Engineers, 39, 539-550.