대출 상환 예측을 위한 의사결정나무모델과 TabNet 간 성능 비교

한수진 [1], 김현철 [2]
[1] 고려대학교 컴퓨터정보통신대학원 빅데이터융합학과 석사과정
[2] 고려대학교 컴퓨터학과 교수

hansujin9@korea.ac.kr, harrykim@korea.ac.kr

Performance comparison between Decision tree model and TabNet for loan repayment prediction

Sujin Han[1], Hyeoncheol Kim[2]
Dept. of Computer & Information Technology, Korea University

요 약

본 연구는 은행에서 리스크 관리 자동화를 위해 고객의 대출 상환 여부 예측 모델을 제안하고자 한다. 예측 모델로 금융 데이터 같은 정형데이터에서 전통적으로 높은 성능을 보인 의사결정나무기반 모델 LightGBM, CatBoost, XGB 와 최근 제안된 정형데이터에서 사용할 수 있는 설명 가능한 딥러닝 기반 모델 TabNet 간의 성능 비교를 진행한다. 다만, 대출 상환 여부 데이터는 불균형 클래스 데이터로 구성되어있어 샘플링을 진행한다. SMOTE, Random Under Sampling, 혼합 방식을 비교해 가장 높은 성능의 샘플링 기법을 제안한다. 대출 상환 여부 예측 결과 TabNet 모델이 의사결정나무 모델들보다 좋은 성능을 보여 정형데이터에서 의사결정나무 기반 모델을 딥러닝 모델이 대체 할 수 있는 가능성을 확인했다.

1. 서론

NPL(Non-Performing Loans)는 3 개월 이상 연체 중인 대출 채권으로 사실상 받을 수 없는 돈으로 여겨진다. 전체 대출 총액에서 NPL 이 차지하는 비중인 NPL 비율은 대표적인 은행 자산 건전성 지표로 통한다. 따라서, 대출 심사 단계에서부터 고객의 대출 상환 여부를 정확하게 파악해 리스크 관리를 자동화 해주는 머신러닝 모델의 역할이 더욱 중요해지고 있는 시점에 본 연구에서는 고객의 대출 상환 여부를 예측해보고자 한다.

대출 상환 예측 데이터에서 대출을 미상환한 고객은 대출을 상환한 고객 클래스 보다 상대적으로 소수이므로 학습시 대출 상환 고객 데이터에 대해 과적합될 위험성이 존재한다. 따라서, 데이터 샘플링 과정에서 불균형 데이터 처리방법인 SMOTE[1], Random Under Sampling, 혼합 방법을 각 예측 모델에 적용됐을 때 평균 성능을 비교해 가장 좋은 샘플링 방법을 채택하고자 한다.

대출 상환 예측 데이터는 정형 데이터로 구성되어 있으며, 금융 산업에서는 금융 감독과 규제 이슈로 실무에서는 해석 가능한 모델에 대한 필요성이 높다. 따라서, 현업에서는 정형데이터에 의사결정나무 기반 모델을 활용하는데 그치고 있다. 그러나, 최근 금융데이터가 고차원의 큰 규모로 생산되며 비정형데이터 및 다양한 연구 분야에서 높은 성능을 보이고 있는 딥러닝을 적용하고자 한다. 딥러닝은 설명이 불가능하며 정형 데이터에서는 다소 낮은 성능으로 실무에 적용하는데 한계가 존재했다. 이후, 의사결정나무를 모방하여 정형데이터에서도 높은 성능을 보이며 해석 가능한 딥러닝 모델이라는 강점을 가지고 있는 TabNet[2]이 등장했다.

따라서, 예측 모델에서 전통적인 강세를 보이는 의사결정나무 기반 모델들과 딥러닝 기반 TabNet 모델 간 성능 비교를 통해 높은 성능의 대출 예측 모델을 제안하고자 한다.

2. 데이터 수집 및 샘플링

본 연구에서는 kaggle 에서 제공하는 22 년 대출 상환 예측 데이터 셋[3]을 사용했다. 12 개의 컬럼과 25 만개의 고객의 데모정보, 소득 수준 등으로 구성된 데이터다. 범주형 속성 컬럼은 7 개, 수치형 속

성 컬럼은 5 개로 구성되어 있으며 클래스의 비율이 9%로 굉장히 불균형한 데이터다.

불균형한 데이터셋으로 대출 상환 고객 데이터에 과적합되어 모델 성능이 하락하는 것을 방지하기 위해 데이터 샘플링을 진행한다. 소수 클래스를 서로 보간해 인공적인 추가 데이터를 만드는 오버샘플링 기법인 SMOTE, 랜덤으로 다수 클래스 데이터를 삭제하는 언더샘플링 방식인 Random Under Sampling, 비율을 설정해 일부 다수 클래스를 언더샘플링 한 후 SMOTE 로 오버샘플링하는 혼합 샘플링 방법까지 3 가지 샘플링 방법론들과 원데이터를 예측 모델에 적용해 성능을 비교했다. 성능 비교는 예측 모델 성능 비교에서 많이 사용하는 F2-score 과 Accuracy, AUC 를 사용해 비교한다. F2-score 을 선택한 이유는 은행은 대출을 미상환할 고객에게 대출이 실행되는 것 더욱 큰 리스크이므로 정밀도와 재현율의 조화 평균인 F1-score 에서 재현율 값에 가중치를 두었다.

3. 대출 상환 예측 모델 성능 비교

대출 상환 예측 모델은 기존 정형 데이터 예측에서 강세를 보이는 의사결정나무 기반 모델 3 가지, 딥러닝 기반 모델 1 가지, 의사결정나무 기반 모델과 딥러닝 기반 모델의 앙상블 모델 3 가지 총 7 가지 모델의 성능을 비교 분석한다.

의사결정나무 기반 모델들은 메모리를 적게 사용하며, 예측을 잘 수행해 다양한 분야의 예측모델에서 높은 성능을 보여왔다. 각 모델 별 특징으로는 LightGBM[4]은 Leaf-wise 방식으로 과적합에 취약하지만 빠른 성능을 보여주며, CatBoost[5]는 범주형 데이터에 특화된 부스팅 알고리즘으로, 범주형 변수를 자동으로 처리한다. XGB[6]는 Level-wise 방식으로 빠른 속도와 높은 정확도를 제공하는 모델이다.

딥러닝 모델 TabNet 은 의사결정나무 모델의 변수 설명력을 가진 딥러닝 모델이다. 딥러닝의 장점은 이미 학습된 모델에 데이터를 추가해 지속적인 학습이 가능하다. TabNet 은 DNN 블록에서 마스킹하지 않은 변수만 학습하도록 해 의사결정나무와 유사한 역할을 한다. 또한, attentive transformer 블록에서 사전 정보량을 집계해 주요 변수를 선택할 수 있어 최종 결과로 해석 가능한 변수 정보를 도출한다.

그리고, 각 LighGBM, CatBoost,XGB 을 TabNet 과 앙상블했을 경우 모델 성능의 변화를 분석한다.

4. 실험 결과

데이터 샘플링 방식은 F2-score 과 AUC 에서 혼합 방식이 가장 좋은 성능을 보이고 그 다음 SMOTE 가 좋은 성능을 보였다. 즉, 데이터가 불균형할 경우 오버 샘플링을 통해 성능을 높이는 것은 유의미한 결과를 보인다.

데이터 예측 모델은 F2-scorer 과 AUC 에서 TabNet 이 모든 의사결정나무 모델들보다 더 좋은 성능을 보였으며, TabNet 다음으로는 의사결정나무 모델들 중에서는 CatBoost 가 좋은 성능을 보였다.

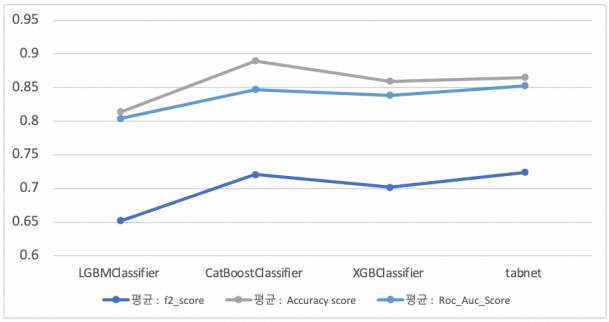
Accuracy 도 CatBoost 를 제외하고는 TabNet 이 전반적으로 더 좋은 성능을 보였다 . Accuracy 는 전체 예측 중 올바른 예측의 비율을 나타내므로, 불균형한 데이터셋에서는 실제 성능을 정확히 반영하지 못할 수 있어 F2_scroe 과 AUC 와 다소 다른 결과를 보일 수 있다.

추가적으로, 의사결정나무 모델들과 딥러닝 모델 간 앙상블을 진행하면 모두 의사결정나무 모델 단독 성능 대비 향상된 성능을 보였다. 최종적으로 CatBoost 와 TabNet 의 앙상블 모델이 가장 좋은 성과를 보였다.

<표 1> 데이터 샘플링 및 대출 상환 예측 모델 결과

데이터 샘플링	모델 구분	F2_score	Accuracy	AUC
Nothing	TabNet	0.705	0.838	0.843
	LightGBM	0.032	0.880	0.512
	LightGBM_ens	0.579	0.892	0.760
	CatBoost	0.584	**0.905**	0.763
	CatBoost_ens	0.714	0.886	0.843
	XGB	0.233	0.889	0.593
	XGB_ens	0.660	0.891	0.808
SMOTE	TabNet	0.681	0.843	0.826
	LightGBM	0.581	0.746	0.756
	LightGBM_ens	0.717	0.871	0.847
	CatBoost	0.715	0.896	0.843
	CatBoost_ens	0.720	0.884	0.847
	XGB	0.681	0.843	0.826
	XGB_ens	0.722	0.876	0.850
Random Under Sampling	TabNet	0.703	0.846	0.841
	LightGBM	0.034	0.880	0.513
	LightGBM_ens	0.576	0.892	0.758
	CatBoost	0.580	**0.903**	0.761
	CatBoost_ens	0.712	0.888	0.841
	XGB	0.233	0.889	0.593
	XGB_ens	0.657	0.893	0.806
Random Under Sampling +SMOTE	TabNet	**0.724**	0.865	**0.853**
	LightGBM	0.583	0.758	0.758
	LightGBM_ens	0.721	0.870	0.850
	CatBoost	0.716	0.896	0.843
	CatBoost_ens	**0.726**	0.884	**0.851**
	XGB	0.681	0.843	0.826
	XGB_ens	0.723	0.876	**0.851**

<그림 1> 대출 상환 예측 모델 평균 성능 그래프



5. 결론

본 연구에서는 대출 상환 가능 여부를 예측하기 위해 불균형한 데이터를 샘플링한 후 의사결정나무 기반 모델과 딥러닝 기반 모델의 성능을 비교했다.

그 결과, 전반적인 성과 지표에서 TabNet 이 의사결정나무 모델들보다 좋은 성과를 보여줬다. 의사결정나무 모델들 간에는 CatBoost, XGB, LightGBM 순으로 좋은 성능 보였다. 이를 통해, 딥러닝 기반 모델도 기존 의사결정나무 기반 모델에 대안으로 사용될 수 있는 가능성을 확인했다. 최종적으로는 CatBoost 와 TabNet 의 앙상블 모델이 가장 좋은 성능을 보였다. 금융업에서도 대용량 데이터 셋이 많이 생성되고 있는 시점에 지속적인 학습이 가능하고 대용량 데이터에서 높은 성능을 보이는 딥러닝 기반 모델이 앞으로도 더욱 좋은 성과를 거둘 것으로 기대된다.

향후 연구로는 다양한 사이즈의 대출 예측 데이터들에서 예측 모델 성능을 비교해, 데이터 크기별 모델 성능이 본 연구와 유사한 결과가 나오는지 추가적으로 확인해 금융 데이터에서 딥러닝 모델의 활용 가능성을 추가 검증하고자 한다.

참고문헌

[1] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer, SMOTE: Synthetic Minority Over-Sampling Technique, Journal of Artificial Intelligence Research, Vol. 16, pp. 321-357, 2002

[2] SercanO, Arık, TomasPfister, TabNet: Attentive Interpretable Tabular Learning, AAAI, Vol. 35, pp.6679-6687, 2020

[3] https://www.kaggle.com/datasets/subhamjain/loan-prediction-based-on-customer-behavior

[4] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, Q. Ye, and T.-Y. Liu., LightGBM: A Highly Efficient Gradient Boosting Decision Tree. Advances in Neural Information Processing Systems, NIPS, Vol.30, pp.3149-3157, 2017

[5] Liudmila Prokhorenkova, Gleb Gusev, Aleksandr Vorobev, Anna Veronika Dorogush, Andrey Gulin, CatBoost: unbiased boosting with categorical features, Advances in Neural Information Processing Systems, pp.6638-6648, 2018

[6] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, SIGKDD , pp.785-794, 2016.