데이터 증강 기법의 앙상블을 통한 레이블 불균형 해소: 설명 가능한 신용평가 모델을 중심으로

정지영 [1], 이소연 [2], 용예린 [3], 김민준 [4]
[1](주)엠로
[2] 숙명여자대학교 경영학과 석사과정
[3] 서강대학교 경제학과 학부생
[4] 한양대학교 경제금융학과 학부생

young991003@naver.com, soyeonri@sookmyung.ac.kr, esther1110@sogang.ac.kr, skyhero97@hanyang.ac.kr

Mitigiating Data Imbalance via Ensembled Data Augmentation: An Explainable Credit Scoring Models

Ji-Young Chung[1], So-Yeon Lee [2], Ye-Lin Yong[3] Min-Jun Kim[4]
[1]EMRO, Seoul, Korea
[2]Dept. of Business Administration, Sookmyung Womens University
[3]Dept. of Economics, Sogang University
[4]Dept. of Economics and Finance, Hanyang University

요 약

최근 금융 분야는 예측 모델의 복잡성으로 인한 블랙박스 문제와 금융 규제에 대한 관심이 높아지고 있다. 이에 따라 금융 업계는 신뢰성과 투명성을 강조하며, 특히 신용평가 분야에서 설명 가능한 모델 연구가 활발히 진행되고 있다. 또한, 해당 분야에서 소수 클래스에 대해 충분히 학습하지 못하고 다수 클래스에 과적합 될 수 있는 데이터 불균형 문제 역시 강조되고 있다. 이는 제 2 종 오류(Type 2 Error)를 최소화해야 하는 상황에서 더욱 부각되며, 대출 상환 능력이 낮은 고객을 최대한 식별해야 하는 개인 신용평가 문제에서 매우 중요한 화두로 떠오르고 있다. 본 논문에서는 어텐션 메커니즘을 활용하여 모델의 설명 가능성을 개선하고, 분석 결과를 해석하는 데 도움이 되고자 한다. 더 나아가, SMOTE, GAN, ADASYN 등 총 다섯 가지 데이터 증강 기법을 실험하여, 이를 앙상블 하였을 때 소수 클래스 레이블에 대한 분류 정확도를 크게 개선할 수 있음을 확인하였다.

1. 서론

신용평가는 소비자 신용을 평가하고 대출 기관들에게 도움을 주는 중요한 위험 관리 도구로 활용되고 있다. 최근 금융 분야에서는 인공지능이 급속하게 발전하며, 신용평가 분야에서도 예측력 향상을 위한 다양한 기계학습 기법을 적용하려는 시도가 있다.[1]
그러나 모델의 복잡도가 증가함에 따라 블랙박스 모델로 인한 내부 동작을 설명하는데 어려움이 발생하고 있다. 또한 일반 개인정보 보호법과 동등 대출 기회법과 같은 규제들로 인해 신용평가 분야에서는 설명 가능성이 더욱 중요해지고 있다.[2]
대출 상환 데이터에서 대출 완전 상환 고객의 수보다 채무 불이행을 한 고객의 수가 적은 것이 특징이다. 이전 연구에서는 불균형 데이터를 활용하여 모델에 대한 Receiver Operating Characteristic curve(ROC) 등의 성능 지표를 중점적으로 다루었으며, 소수 클래스에 대한 관심이 부족한 경향이 있었다.[3] 이처럼 기존 다수의 국내외 신용평가 연구에서 정확도, 정밀도 등의 성능 지표에 초점을 맞췄으며, 소수 클래스에 성능에 대한 관심은 부족했다.
Adou and Pointon(2011) 연구에서는 제 1 종 오류를 좋은 신용을 가진 대출자를 신용 불량으로 분류하는 경우로 정의하고, 제 2 종 오류를 신용불량인 대출자를 좋은 신용을 가진 대출자로 잘못 분류한 경우로 정의했다. 제 2 종 오류가 더 큰 비용을 초래하며 비용 비율은 약 5:1 정도라고 언급했다.[4] 따라서 제 2 종 오류에 취약한 개인 신용평가에서 소수 클래스 레이블의 분류 정확도를 개선하는 것이 중요하다.

본 논문은 어텐션 메커니즘을 활용한 설명 가능성을 개선하고 분석 결과를 해석할 수 있는 개인 신용평가 모델을 연구하고자 한다. 더 나아가, 데이터 불균형 문제를 해소하기 위해 Synthetic Minority Over-sampling Technique (SMOTE), Generative Adversarial Networks(GAN) 등의 데이터 증강 기법을 활용하였다. 각 기법의 장점을 활용하고 단점을 보완하기 위해 앙상블 기법을 적용하여 소수 클래스 레이블에 대한 분류 성능을 향상시키는 방법을 제안한다.

2. 문헌 연구

2.1 SMOTE

SMOTE 는 소수 클래스에서 합성된 샘플을 생성하는 방법으로, K-Nearest-Neighbors(KNN)를 활용하여 소수 클래스에 속한 샘플과 이웃한 샘플을 찾은 후 두 샘플 사이의 내분점에 새로운 샘플을 생성한다. 이후 소수 클래스에 속한 각 샘플에 대해 마찬가지로 새로운 샘플 생성을 반복 수행한다.[5]

2.2 Borderline SMOTE

Borderline Synthetic Minority Oversampling Technique (Borderline SMOTE)는 클래스 간의 경계에 존재하는 샘플의 수를 늘림으로써 학습이 어려운 소스 클래스에 주목한다. 이 방법에서는 특정 데이터의 주변에 위치한 데이터 중 과반수 이상이 다중 클래스에 속할 때 이를 경계라고 판단하며 이러한 경계에 존재하는 데이터에 대해서만 SMOTE 기법을 적용한다.[6]

2.3 ADASYN

Adaptive Synthetic Sampling(ADASYN)는 불균형 데이터 분포를 보완하기 위해 가중 분포를 활용하여 합성 데이터를 생성하는 기법이다. 다수 클래스 관측치가 많이 포함된 KNN 영역을 가진 소수 관측치에 더 많은 가중치를 부여하여 더 많은 합성 관측치를 생성한다.[7]

2.4 CTGAN

GAN 모델을 활용해 데이터를 생성할 때, 범주형 데이터와 연속형 데이터가 함께 포함된 경우 문제가 발생할 수 있다. 연속형 데이터의 경우, 확률 분포가 다중 모드 분포를 가질 수 있으며, 범주형 데이터는 각 범주의 빈도수가 다르고 불균형한 성질을 보일 수 있다. 이러한 문제를 해결하기 위해 Conditional Tabular GAN(CTGAN)이 제안되었다.[8]

2.5 SMOTified GAN

SMOTE 기법은 기존 데이터를 단순하게 오버샘플링하기 때문에 데이터의 다양성이 부족할 수 있다. 반면, GAN 은 원 데이터가 지닌 확률분포를 추정하기 때문에 SMOTE 로 생성된 데이터보다 더 현실적이지만 소수 클래스의 데이터 수가 제한적인 경우 단독으로 사용하기 어렵다. 이러한 한계를 극복하기 위해 해당 논문에서 SMOTified GAN 을 제안했다.[9]

3. 데이터 및 제안 모델

3.1 데이터

이 연구는 랜딩 클럽[1]에서 제공하는 개인 대출 신청자 데이터 셋이 사용되었다. 데이터는 2007 년부터 2020 년 3 분기까지 수집되었으며, 종속변수로 ‘Loan Status’를 사용하였다.

이 변수는 'Fully Paid', 'Charged Off', 'Default' 등 다양한 상태를 포함하고 있다. 이 중에서 'Fully Paid', 'Charged Off', ‘Default’는 대출 상환 여부, 'Late (31-120 days)'와 'Late (16-30 days)'는 연체 상태를 나타내는 변수로 사용되었다. 두 연체 상태는 대출 상환 지연을 의미하므로 연체 상태 자체를 파악하기 위해 ‘Late’로 통합하여 사용되었다.

기존 많은 연구에서는 대출 상환 여부에 대한 이진 클래스 분석에 주로 초점을 맞추었으나, 본 논문에서는 Fully Paid, Charge Off, Late, Default 와 같은 다중 클래스를 다룬다.[10, 11] 따라서 종속 변수에 대해서는 Fully Paid(0), Charge Off(1), Late(2), Default(3)으로 변경하여 활용했다. 새로운 변수로는 '대출신청년도' 변수를 생성하였다. 더불어, 미국 경제의 상황을 반영하며, 대출 상환 및 연체에 영향을 미칠 수 있는 외부 요소를 고려하기 위해 연도별 미국의 실질 GDP 데이터를 활용하여 '국면전환' 변수를 추가하였다. 금융 위기 기간인 2008 년도에 금융 시장에 중요한 특성을 나타낼 가능성이 높기 때문에 2008 년의 결측치 비율이 100%인 변수를 제거했다. 또한, 다중공선성이 10 이상인 변수와 결측값이 있는 행은 모두 제거하였다. 범주형 변수의 경우 라벨 인코딩과 원-핫 인코딩을 시행하였고 수치형 변수에 대해서는 min-max scaler 를 적용하여 모두 43 개의 변수를 예측 모델에 사용하였다.

대출 신청 데이터는 과거의 대출 이력을 기반으로 미래의 대출 상환 여부를 예측하는 시계열의 특성을 가지고 있다. 이에 따라 2007 년부터 2018 년까지의 데이터를 훈련 데이터로, 2019 년부터 2020 년 3 분기까지의 데이터를 테스트 데이터로 나누었다.

[1] 랜딩 클럽 데이터 출처: https://www.lendingclub.com/

3.2 제안 모델

어텐션 메커니즘은 주로 시퀀스 투 시퀀스 작업에서 주로 사용되며, 입력 시퀀스의 각 요소가 출력 시퀀스의 어떤 부분에 초점을 맞춰야 하는지를 학습하는 데 도움을 준다.[12] 인코더의 은닉 상태와 디코더의 현재 은닉 상태 간의 관련성을 계산하여 어텐션 스코어를 얻고, 어텐션 스코어는 확률분포로 변환되어 인코더의 각 은닉 상태의 중요도를 나타낸다. 이를 통해 모델은 입력 시퀀스의 다양한 부분에 주의를 기울여 출력 시퀀스를 생성한다.

대출 상환 예측 모델에서 어텐션 메커니즘은 각 변수의 중요도를 고려하여 설명력을 제공한다. 변수의 가중치와 입력 벡터를 행렬 곱셈을 통해 계산하여 대출의 상환 여부를 예측하는 모델을 개발하였다.

4. 실험 결과

본 연구에서는 다섯 가지의 데이터 증강 기법인 SMOTE, Borderline SMOTE, ADASYN, CTGAN, SMOTified GAN 을 사용한 데이터 증강 기법의 앙상블에 대한 성능을 비교하고자 한다.

<표 1> 데이터 증강 기법의 앙상블을 적용

Attention - Layer 1	label#0	label#1	label#2	label#3	mean	w-mean
smotegan + ctgan + ADASYN	0.9209	0.8859	0.6436	0.5072	0.7394	**0.6721**
smote	0.9216	0.8661	0.8578	0.2899	0.7339	**0.6671**
BorderlineSMOTE + smote + ADASYN	0.7635	0.9303	0.6328	0.5217	0.7121	**0.6665**
ADASYN + ctgan + smote	0.8639	0.9463	0.5213	0.5652	0.7242	0.6559
BorderlineSMOTE + ctgan + smote	0.8949	0.8478	0.5792	0.5362	0.7145	0.6495
smotegan + ctgan + smote	0.8919	0.7522	0.7530	0.3768	0.6935	0.6351
ADASYN	0.9907	0.6850	0.9204	0.1884	0.6961	0.6242
BorderlineSMOTE + ctgan + ADASYN	0.8891	0.8438	0.4219	0.6232	0.6945	0.6235
ADASYN + smotegan + smote	0.8668	0.6930	0.8450	0.2754	0.6701	0.6174
BorderlineSMOTE + smotegan + ADASYN	0.8776	0.7560	0.9109	0.1594	0.6760	0.6136
smotegan + smote	0.9338	0.8142	0.7706	0.2464	0.6913	0.6122
BorderlineSMOTE + smotegan + smote	0.8707	0.7141	0.8997	0.1884	0.6682	0.6107
BorderlineSMOTE	0.9801	0.7695	0.9470	0.0725	0.6923	0.6087
No Augment	0.9937	0.7916	0.9886	0.0000	0.6935	0.6037
smotegan + ctgan	0.8508	0.8783	0.9484	0.0290	0.6766	0.6028
smotegan + ctgan + BorderlineSMOTE	0.9664	0.7340	0.8307	0.1304	0.6654	0.5798
smote + BorderlineSMOTE	0.2171	0.9558	0.9963	0.0000	0.5423	0.5616
ctgan	0.9886	0.5179	0.9859	0.0145	0.6267	0.5526
ctgan + smote	0.2382	0.9570	0.0414	0.8841	0.5302	0.5391
ctgan + ADASYN	0.0620	0.9863	0.0236	0.9130	0.4962	0.5313
ADASYN + smotegan	0.9254	0.2913	0.7118	0.3333	0.5655	0.5166
ADASYN + BorderlineSMOTE	1.0000	0.2628	0.9952	0.0145	0.5681	0.5059
ADASYN + smote	1.0000	0.2612	0.9957	0.0145	0.5678	0.5058
smotegan + BorderlineSMOTE	0.9859	0.6596	0.6368	0.0870	0.5923	0.4838
ctgan + BorderlineSMOTE	0.2190	0.1090	0.9767	0.2284	0.3833	0.4655
smotegan	0.0000	0.9995	0.4523	0.0000	0.3630	0.3582

<표 1>은 다섯 가지 데이터 증강 기법과 원본 데이터 셋에 대한 레이블 정확도 및 가중 평균을 비교한 결과이다. Late 와 Default 는 대출 이행의 초기 단계에서 주요한 역할을 하며, 재정적 영향 등 다양한 측면에서 중요하다. 따라서 레이블(0)은 0.1, 레이블(1)은 0.2, 레이블(3)과 레이블(4)은 0.35 의 가중치를 부여했다. 3 개까지 데이터 증강 기법을 앙상블 하여 모델에 적용한 결과, SMOTified GAN + CTGAN + ADASYN 조합은 가중 평균 정확도가 0.6721 이고, 평균 정확도 역시 0.7394 로 가장 우수한 성능을 보였다. SMOTifed GAN 은 레이블 1 에서 다른 기법보다 우수한 성능을 보였으므로 레이블 1 의 성능이 향상된 것으로 예상한다. CTGAN 과 ADASYN 을 개별적으로 사용했을 때 레이블 0 과 2 에서 높은 성능을 나타내었으며, 이를 조합하면서 더 다양한 데이터 패턴을 학습한 것으로 보인다. ADASYN 의 레이블 3 의 성능은 크게 높지는 않았지만, 다양한 데이터 증강 기법을 조합하면서 다양한 패턴을 학습하면서 전반적인 성능이 향상된 것으로 예상한다.

3 개의 증강 기법을 앙상블 한 결과, 가중평균 정확도가 대체로 향상되었으며, 특히 ADASYN 기법이 포함된 경우 성능이 향상된 것으로 관찰되었다.

<표 2> 상위 3 가지 앙상블 기법의 성능 비교

Attention - Layer 1		label#0	label#1	label#2	label#3	Accuracy	Precision	Recall	f1-score
smotegan + ctgan + ADASYN	Test only	0.9209	0.8859	**0.6436**	**0.5072**	0.8975	0.6300	**0.7394**	0.6244
	Augment only	0.8791	0.5012	0.3408	0.6172	0.5846	0.6052	0.5846	0.5806
	Train split	0.0000	0.9625	0.9779	0.0000	0.1925	0.3646	0.4851	0.1572
smote	Test only	0.9216	0.8661	0.8578	0.2899	0.9084	0.6187	0.7338	0.6408
	Augment only	0.9681	0.8417	0.8988	0.2864	0.7488	0.7699	0.7488	0.7310
	Train split	0.9674	0.8827	0.8083	0.2209	0.9500	0.7015	0.7198	0.6879
BorderlineSMOTE + smote + ADASYN	Test only	0.7635	0.9303	0.6328	0.5217	0.7810	0.6102	0.7121	0.5751
	Augment only	0.7617	0.9519	0.6401	0.5714	0.7313	0.7443	0.7313	0.7338
	Train split	0.7978	0.9700	0.5839	0.5930	0.8305	0.621	0.7362	0.5808

<표 2>는 가중평균 정확도가 높은 상위 3 개의 기법의 정확도, 정밀도, 재현율, F1 스코어 성능을 비교한 결과이다. 실험에서는 테스트 데이터를 세 가지 방식으로 나누어 성능을 비교했다.

(1) Test Only 는 시간을 기준으로 분리된 테스트 셋에서 모델이 얼마나 잘 일반화하는지를 평가한다.

(2) Train Split 은 시간과 무관하게 랜덤 분할된 셋에서 얼마나 일반화하는지를 보기 위함이다. 만약 모델이 Train Split 에서 높은 성능을 보이지만 Test Only 에서 낮은 성능을 보인다면, 모델의 성능이 시간 변수나 매크로 경제 상황에 영향을 받을 수 있음을 시사한다.

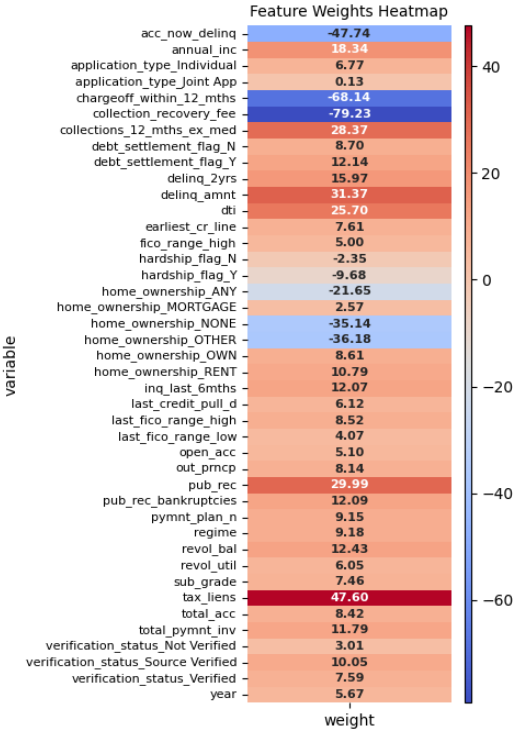
(3) Augment Only 는 데이터 증강 기법을 활용하여 생성한 테스트 셋을 사용한다. 모델이 Augment Only 에서만 우수한 성능을 보이고 기타 데이터 셋에는 성능이 낮다면, 데이터 증강에 사용한 증강 기법이 과하게 과적합 되었을 가능성을 고려할 수 있다.

성능을 비교할 때, 제 2 종 오류가 고려되는 재현율을 주요 비교 지표로 사용했다.

SMOTified GAN + CTGAN + ADASYN 조합에서 Test Only 의 재현율이 0.7394 로 가장 우수한 성능을 보였다. 이는 제 2 종 오류에 취약한 개인 신용평가 모델에서 소수 클래스 레이블의 분류 정확도를 개선하여 오류 비용을 감소시킬 수 있음을 시사한다.

Augment Only 의 경우 재현율은 0.5846 으로 원본데이터와 비교했을 때 성능이 낮았지만, 증강된 데이터가 모델에 어느 정도 영향을 미치는 것을 확인할 수 있었다. Train Split 의 경우, SMOTE 와 Borderline SMOTE + SMOTE + ADASYN 에 비해 낮은 재현율을 보였다. 이 결과는 GAN 기법을 다수 사용할 경우, 시계열 데이터를 고려한 Test Only 에서 더 나은 성능을 기대할 수 있음을 시사한다. 대출 데이터와 같은 시계열 데이터를 다룰 때 GAN 기법을 사용하는 것이 더 유용할 것으로 판단된다.

5. 변수 중요도



<그림 1> Default 에 대한 변수 가중치 시각화

<그림 1>은 신용평가모형을 활용하여 Default 변수에 대한 변수 가중치를 시각화한 것이다. 이전 대출에서 미수금을 회수하기 위해 지불한 수수료(collection_recovery_fee), 최근 12 개월 동안 채무 불이행 수(chargeoff_within_12_mths)는 Default 예측에 큰 영향을 미친다. 또한, 이전 대출에서 미수금을 회수하기 위해 지불한 수수료는 음의 가중치를 가지며 대출 상환 여부에 부정적인 영향을 준다는 것을 나타낸다.

6. 결론 및 향후 연구 과제

본 연구는 랜딩 클럽 데이터를 활용하여 데이터 증강 기법의 앙상블과 어텐션 메커니즘을 사용하여 분석결과를 해석하는데 도움을 주었다. 실험 결과, 어텐션 메커니즘을 통해 신뢰성이 중요한 금융 분야에서 보다 유용하게 활용할 수 있음을 확인했다. 또한, 개별 증강 기법보다 앙상블을 활용함으로써 불균형한 레이블 클래스에 대한 성능을 향상시킬 수 있음을 확인하였다. 특히 GAN 기법을 활용한 결과가 시계열 데이터에 효과적으로 적합하며 우수한 성능을 보였다.
향후 연구 방향으로는 제안된 앙상블 방법이 다른 모델에서도 효과적으로 동작하는지에 대한 추가 연구가 필요하다. 또한, GAN 기법을 사용하지 않은 경우 Train Split 에서 더 나은 성능을 관찰한 점을 고려하여 원인을 분석하고, 경제 외부 요인을 고려하는 등 추가적인 연구가 필요하다.

참고문헌

[1] 천예은, 김세빈, 이자윤, 우지환, “설명 가능한 AI 기술을 활용한 신용평가 모형에 대한 연구.” 한국데이터정보과학회지, 제 32 권, 제 2 호, 283-295, 2021

[2] Kong, Y., Wang Y., Sun S., and Wang J., “XGB and SHAP credit scoring model based on Bayesian optimization”, Computing and Electronic Information Management, Vol. 10, No. 1, 46–53, 2023

[3] Chen, D., Ye W., and Ye J., “Interpretable selective learning in credit risk”, arXiv:2209.10127, 2022

[4] Abdou, H., and Pointon, J. “Credit Scoring, Statistical Techniques and Evaluation Criteria: A Review of the Literature.”, Intelligent Systems in Accounting, Finance & Management, 18, 2-3, 59-88, 2011

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. “SMOTE: Synthetic Minority Over-sampling Technique”, arXiv:1106.1813, 2011

[6] Han, H., Wang, W.-Y., and Mao, B.-H. “Borderline-smote: A new over-sampling method in imbalanced data sets learning”, LNCS, 3644, 878-887, 2005

[7] He, H., Bai, Y., Garcia, E. A., and Li, S., “ADASYN: Adaptive synthetic sampling approach for imbalanced learning”, IEEE world congress on computational intelligence, 1322-1328, 2008

[8] Xu, L., Skoularidou, M., Cuesta-Infante, A., and Veeramachaneni, K. “Modeling Tabular data using Conditional GAN”, arXiv:1907.00503, 2019

[9] Sharma, A., Singh, P. K., & Chandra, R., “SMOTified-GAN for class imbalanced pattern classification problems”, Vol 10, 30655 – 30665, 2022

[10] Chen, Y., Calabrese, R., and Martin-Barragan , B., “Interpretable machine learning for imbalanced credit scoring datasets”, European Journal of Operational Research, 312, 357–372, 2023

[11] J. Xiao, Y. Wang, J. and Chen et al., “Impact of resampling methods and classification models on the imbalanced credit scoring problems”, Information Sciences, 508–526, 2021

[12] Niu, Z., Zhong, G., and Yu, H., “A review on the attention mechanism of deep learning”, Neurocomputing, 48–62, 2021