기업평가 부실예측 머신러닝활용사례 분석과제

머신러닝 알고리즘 활용 논문 분석 및 실무사례

'기업 부도모형의 기계학습 적용과정에 대한 연구' (저자:조동유)

CONTENTS

논문의 연구 배경 및 목적

연구 배경

- 1997년 외환위기와 2008년 글로벌 금융위기, 그리고 COVID-19 팬데믹으로 인해 한국 기업의 부도 확률이 크게 증가함
- 금융위기의 반복적인 발생으로 인해 기업 부도를 사전에 예측하기 위한 연구가 지속적으로 필요하게 됨
- 여신금융기관의 신용위험 관리와 채무불이행 예측의 중요성 강조

연구 중요성

- 기존 연구는 주로 기계학습 기법 간의 모형 성능 비교에 초점을 맞췄음으나 데이터 특성을 정확하게 반영하고 검증하는 과정의 중요성이 강조됨.
- 불균형 데이터 해결의 필요성
  불균형 데이터로 인해 발생하는 예측력 저하 문제를 해결하는 것이 중요함

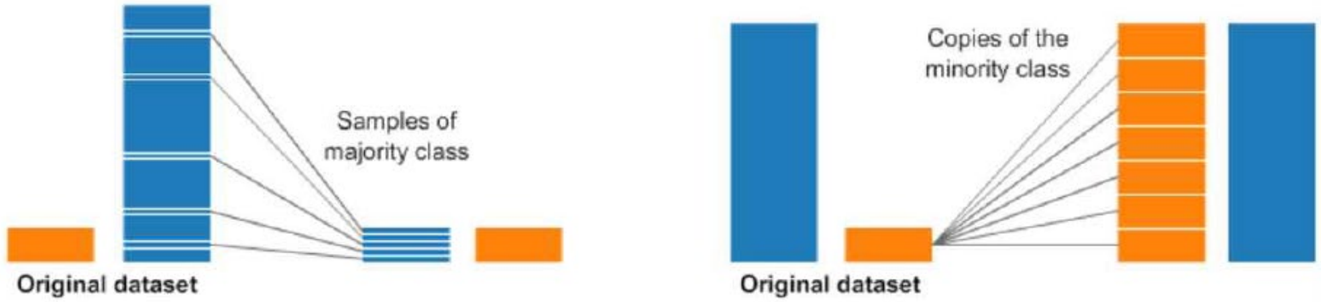
연구의 목적

- 데이터의 불균형 및 시계열적 특성이 모형 예측력에 미치는 영향을 파악하고 부도모형을 구축하는 합리적인 기준을 제시
- 모형 예측력 향상
  특히 oversampling 기법이 어떻게 모형의 예측력을 향상시키는지에 중점을 두고 연구를 진행

논문연구 이론적 배경

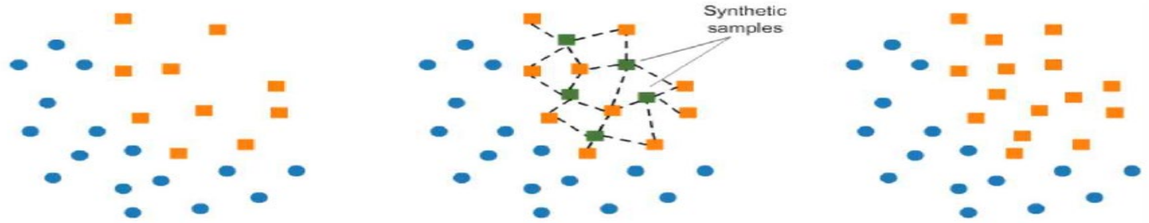
데이터 불균형 해결 전략 분석

불균형 및 시계열 데이터 처리 기법



(a) undersampling (b) oversampling

- undersampling 기법의 한계: 다수범주의 데이터 삭제로 인한 정보 손실과 예측력 변동성이 발생할 수 있다.
- oversampling 기법의 효과: 소수범주 데이터를 증가시켜 다수범주와 비율을 맞추어 예측력을 향상시킨다.
- undersampling 기법은 다수범주의 데이터를 삭제해 정보의 손실을 초래 하며 이러한 문제를 해결하기 위해 oversampling 기법이 더 널리 사용 됨



<그림 2> SMOTE

oversampling 기법에는 random oversampling(ROS)과 synthetic minority oversampling technique(SMOTE) 등이 있습니다. ROS는 기존 소수범주 데이터를 단순 복제하여 데이터를 증가시키는 방법입니다. 반면, SMOTE는 소수범주 사이에 새로운 데이터를 생성해 다수범주와 비율을 맞추는 방법

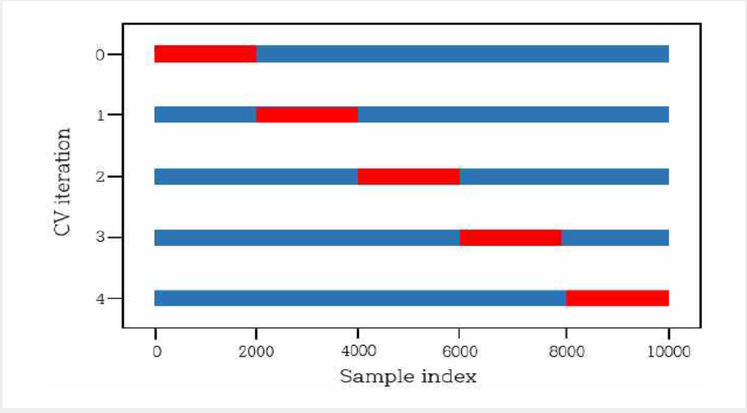
이론적 배경

교차검증(Cross Validation)

기계학습 모델의 성능을 검증하기 위해 학습데이터와 시험데이터로 나눈 후, 학습데이터를 다시 학습세트와 검증세트로 나누어 반복적으로 학습과 검증 과정을 진행하는 방법

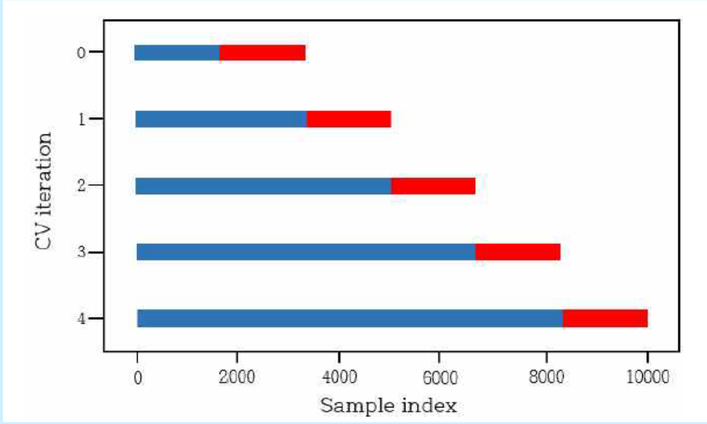
k-fold 교차검증(k-fold cross validation)

- k-fold 교차검증 방법: 학습데이터를 k개의 세트로 나누어 각각 다른 검증세트를 선택하여 k번 반복 검증.
- hyperparameter 최적화: 최적의 hyperparameter를 선택하여 성능이 가장 좋은 모형을 적합.
- 시간 순서 무작위 선택: 일반적으로 시간 순서에 상관없이 무작위로 검증세트를 선택하여 횡단면 데이터에 적합.
- 시계열 데이터의 문제: 시계열 데이터에 k-fold 교차검증을 적용하면 미래 데이터로 학습한 후 과거 데이터로 검증하는 오류 발생.



전진교차검증(Foward Cross-Validation)

- 전진교차검증의 목적 : k-fold 교차검증의 단점을 보완하기 위해 제안된 방법으로, 특히 시계열 데이터를 이용한 모형 적합에 사용됨.
- 검증 세트 선택 방법 : 무작위로 검증 세트를 선택하는 k-fold 교차검증과 달리, 시간의 순서대로 검증 세트를 선택.
- 학습 및 검증 과정 : 시간 순서에 따라 학습 세트를 사용해 모형을 학습하고, 이후 검증 세트를 통해 모형 성능을 검증.
- 적용 분야 : 시계열 데이터를 사용하는 기계학습 모형 적합에 많이 사용되며, 부도예측모형에서도 과거 데이터를 통해 미래 부도를 예측하는 데 적합.



연구 절차 및 방법론

분석 절차 개요:

- 부도 정의 및 설명변수 선택: 선행연구를 참고하여 부도 정의와 설명변수를 선택.
- 데이터 전처리: 결측치 보정, 이상치 조정, 표준화 등의 전처리 진행.

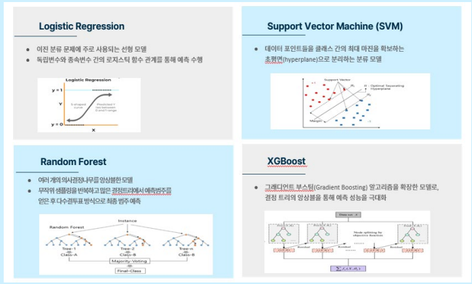
모형 성능 비교:

1. Oversampling 기법 적용 효과 확인:
   - 원데이터, ROS 기법, SMOTE 기법을 통해 데이터 불균형 해소.
   - 각 데이터를 사용하여 기계학습 기법에 적합 후 성능 비교.

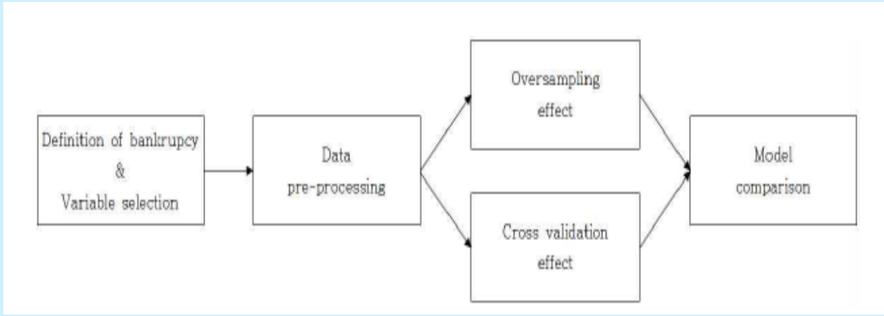
2. 교차검증 효과 확인:
   - k-fold 교차검증, 전진 교차검증 기법 적용
   - 기계학습 기법에 적합 후 모형 성능 비교

3. Oversampling 기법 적용 시점 비교:
   - Oversampling 기법 적용 시점을 달리하여 모형 성능 비교.

- 사용된 기계학습 기법



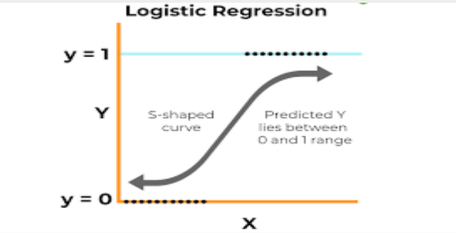
- Logistic Regression
- Support Vector Machine (SVM)
- Random Forest
- Extreme Gradient Boosting (XGBoost)



본연구에 활용한 머신러닝 모델

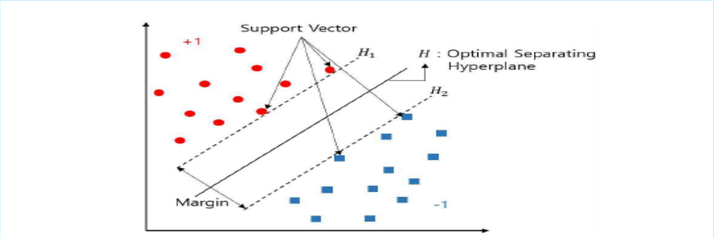
Logistic Regression

- 이진 분류 문제에 주로 사용되는 선형 모델
- 독립변수와 종속변수 간의 로지스틱 함수 관계를 통해 예측 수행



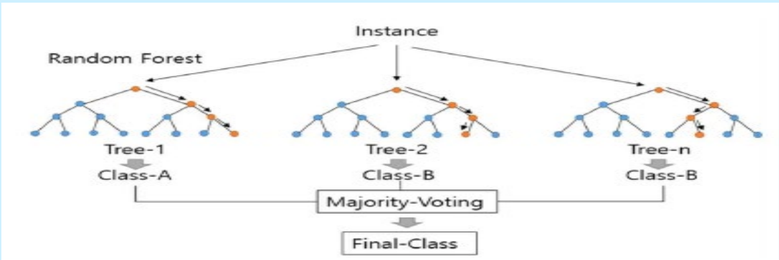
Support Vector Machine (SVM)

- 데이터 포인트들을 클래스 간의 최대 마진을 확보하는 초평면(hyperplane)으로 분리하는 분류 모델



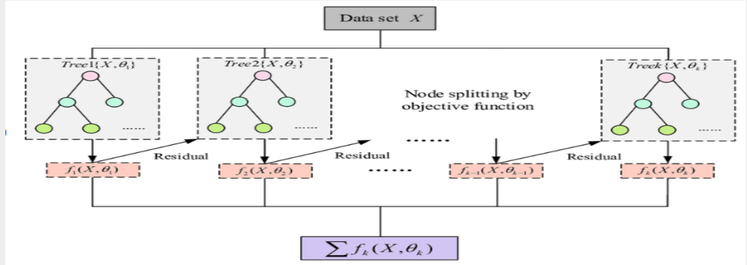
Random Forest

- 여러 개의 의사결정나무를 앙상블한 모델
- 무작위 샘플링을 반복하고 많은 결정트리에서 예측범주를 얻은 후 다수결투표 방식으로 최종 범주 예측



XGBoost

- 그래디언트 부스팅(Gradient Boosting) 알고리즘을 확장한 모델로, 결정 트리의 앙상블을 통해 예측 성능을 극대화

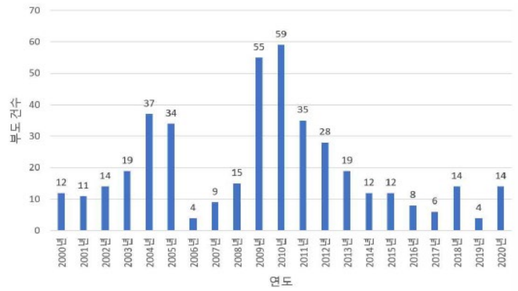


활용데이터 및 변수

활용데이터

- 대상: 한국거래소 코스피 및 코스닥 시장에 상장된 기업 (2,497개)
- 2000년부터 2020년까지 재무 및 부도정보

* 부도의 기준: 파산 신청, 사업체 폐쇄, 채무 조정 신청, 90일 이상 연체 등



부도위험 변수

재무변수/ 파생재무변수

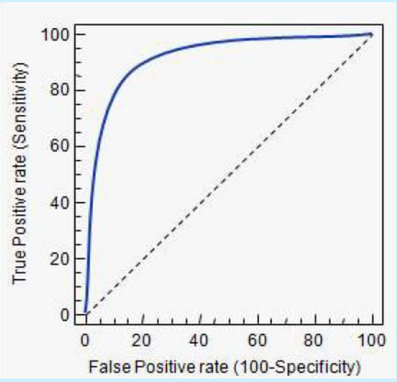
연도별 부도 기업 수:
2008년 글로벌 금융위기 이후 증가,
2008–2013년 사이에 부도 사건 211건

- 데이터 분할: 학습데이터(2000–2009), 시험데이터(2010–2020)
- 학습데이터: 정상기업 48,295개, 부도 기업 210개.
- 시험데이터: 정상기업 71,952개, 부도 기업 211개

<표 4> 최종 설명변수

종류	시점	변수
재무변수	0	CLTL, lnTA
	1	WCTA, SLTA
	2	NITA, CASHTA
	3	EQTA, TLTA, CLTA, FATA
	4	TLEQ, CLCA, CACL, RETA, EBTA
증감율 변수	0	총자산, 총부채, 총자본, 당기순이익, 매입채무, EQTA, NITA
	1	세전계속사업이익, 매출액, EBTA
	2	유동부채, 비유동자산, CATA, FATA, WCTA
	3	영업고정자산, CLTA, CLTL, CLCA, CACL, lnTA
	4	현금 및 현금성자산, TLTA, CASHTA, SLTA, SLEQ, SLFA

설명변수별 4분기전까지의 과거 값을 구한 후 t-test를 통해 정상데이터와 부도데이터의 차이가 가장 큰시차 값을 선택



성능평가 지표

성능 평가 지표

정오분류표(confusion matrix)를 통해 값을 확인한 후, 정확도(accuracy), 오류율(error rate), AUC(area under the curve), 민감도(sensitivity), 특이도(specificity), 정밀도(precision) 등과 같은 값을 통해 평가

Confusion matrix

- Positive 는 부도 기업을 의미 , Negative 는 정상기업
- □□□ □□□□ □□□ □□□□ □□□□ □□□□□ □□□ □□□□□ □□□ 범할 수 있다 (Burez, Van Den Poel, 2009). 소수집단의 관측치가 현저히 적기 때문에 □□ □□□□ □□□□□ □□□□□ □□□ □□□□ □□□□.

		Predicted Class	
		Negative	Positive
Actual Class	Negative	TN (True Negative)	FP (False Positive)
	Positive	FN (False Negative)	TP (True Positive)

$$accuracy = \frac{TN+TP}{TN+TP+FN+FP}$$

$$error\ \ rate = (1-accuracy) = \frac{FN+FP}{TN+TP+FN+FP}$$

민감도(sensitivity), AUC (Area Under the Curve)

- 소수범주인 부도 기업을 부도로 정확하게 분류하는 비율을 의미하는 □□□□ □□□□ □□□□□ □□□
- 민감도의 값이 높을수록 부도 기업을 부도로 정확하게 분류할 확률이 높다고 할 수 있다 .

$$sensitivity = \frac{TP}{TP+FN}$$

- □□□□ □□□□ □□ □□□ □□□ AUC 성능평가지표
- 민감도와 특이도를 함께 반영하여 정확히 분류된 TP 와 잘못 분류된 FP 의 □□□ □□

평가지표 및 성능

ROS와 SMOTE 두 oversampling□법 □□□ Logistic Regression, □□□ SVM, XGBoost를 □용□ 모형에서는 ROS를 사용하였을 때상대적으로 성능이 □□□□□

Random Forest를 적용한 모형에서는 SMOTE를 □□□□□ 때 □능□ □□□

Hulse etal(2007)의 □□□□ oversampling □법 □□ □□ □□□ □□하□ □□□□ □□□, □□□ □□ □절□ oversampling □□□ □□□□□ □다□ □□□ □관□ □□이□

Oversampling 유무에 따른 모형 성능

불균형 데이터에서는 정확도뿐만 아니라 민감도, 특이도, 정밀도, AUC 등의 지표를 함께 고려하여 평가

		Accuracy	AUC	Sensitivity
Logistic Regression	Raw data	0.9969	0.5094	0.0190
	ROS	0.9282	0.8624	0.7962
	SMOTE	0.9358	0.8426	0.7488
Linear SVM	Raw data	0.9971	0.5000	0
	ROS	0.9325	0.8693	0.8057
	SMOTE	0.9418	0.8692	0.7962
Non-linear SVM	Raw data	0.9971	0.5000	0
	ROS	0.9087	0.8691	0.8294
	SMOTE	0.9356	0.8283	0.7204
Random Forest	Raw data	0.9971	0.5047	0.0095
	ROS	0.9569	0.8035	0.6493
	SMOTE	0.9382	0.8367	0.7346
XGBoost	Raw data	0.9969	0.5472	0.0948
	ROS	0.9169	0.8497	0.7820
	SMOTE	0.9248	0.8300	0.7346

원데이터 사용의 한계: 데이터 불균형 해소 없이 모형 구축 시 부도 기업 예측 성능이 낮음
oversampling 기법 필요성: 불균형 데이터 사용 시 oversampling 기법 적용 후 모형 구축 시 성능 개선 확인

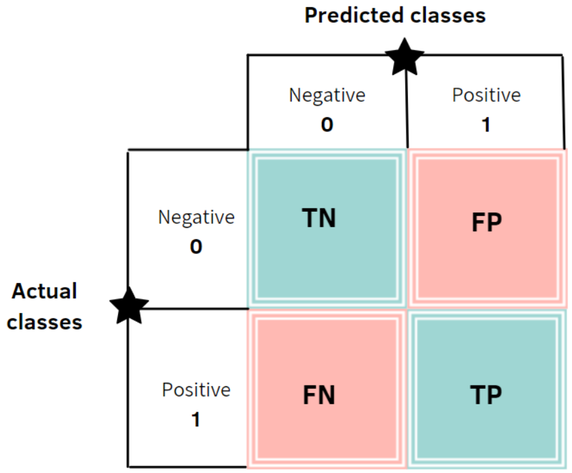
원데이터 사용 시 성능

- 모든 기계학습 모형의 정확도는 99% 이상으로 매우 높음.
- 그러나 대부분 정상기업으로 분류, 부도 기업 예측 성능은 매우 낮음.
- AUC 약 50%, 민감도 0%에 가까움.
- 211개의 부도 기업을 제대로 예측하지 못함.

oversampling 기법 적용 시 성능

- 정확도는 상대적으로 낮지만, 부도 기업 예측 성능은 개선됨.
- AUC가 80% 이상으로 부도 기업 예측 성능 대폭 개선.
- 부도 기업을 대부분 정확히 예측

평가지표 및 성능



Oversampling 유무에 따른 Confusion Matrix

불균형 데이터에서는 정확도뿐만 아니라 민감도, 특이도, 정밀도, AUC 등의 지표를 함께 고려하여 평가

		TN	FP	FN	TP
Logistic Regression	Raw data	71938	14	207	4
	ROS	66814	5138	43	168
	SMOTE	67376	4576	53	158
Linear SVM	Raw data	71952	0	211	0
	ROS	67124	4828	41	170
	SMOTE	67794	4158	43	168
Non-linear SVM	Raw data	71952	0	211	0
	ROS	65397	6555	36	175
	SMOTE	67364	4588	59	152
Random Forest	Raw data	71951	1	209	2
	ROS	68913	3039	74	137
	SMOTE	67545	4407	56	155
XGBoost	Raw data	71919	33	191	20
	ROS	66004	5948	46	165
	SMOTE	66584	5368	56	155

원데이터 사용 시 성능

- 모든 기계학습 모형의 정확도는 99% 이상으로 매우 높은 것으로 보임
- 왼쪽 표 정오분류표(Confusion Matrix)에서 확인할 수 있듯이 이 결과는 불균형 데이터에서 대부분을 다수범주로 즉, 정상기업으로 분류하여 나타난 결과임

- 211개의 부도 기업을 제대로 예측하지 못함.
- 소수범주인 부도 기업의 관점에서 모형의 성능을 확인해보면 AUC는 약 50%이며 민감도는 0%에 가깝게 나타나 모형의 성
- 능이 매우 좋지 않다는 것을 확인할

oversampling 기법 적용 시 성능

- 두 가지 oversampling 기법을 적용한 기계학습 모형의 성능은 원데이터를 사용한 모형에 비해 정확도는 상대 적으로 낮은 것으로 확인된다. 하지만 소수범주 즉, 부도 기업에 대한 예측 성능은 원 데이터에 비해 상당히 개선된 것으로 확인된다

평가지표 및 성능

교차검증에 따른 모형 성능 비교

k-fold 교차검증 vs. 전진교차검증

		k-fold cross validation		Forward cross validation	
		AUC	Sensitivity	AUC	Sensitivity
Logistic Regression	ROS	0.8624	0.7962	0.8624	0.7962
	SMOTE	0.8426	0.7488	0.8426	0.7488
Linear SVM	ROS	0.8693	0.8057	0.8693	0.8057
	SMOTE	0.8692	0.7962	0.8624	0.7867
Non-linear SVM	ROS	0.8691	0.8294	0.8337	0.7678
	SMOTE	0.8283	0.7204	0.8283	0.7204
Random Forest	ROS	0.8035	0.6493	0.7662	0.5545
	SMOTE	0.8367	0.7346	0.8376	0.7346
XGBoost	ROS	0.8497	0.7820	0.8497	0.7820
	SMOTE	0.8300	0.7346	0.8300	0.7346

시계열적 특성을 고려하지 않는 k-fold 교차검증을 적용하였을 때 모형의 성능은 우수하였지만, 모형을 추정하는 과정에서 논리적인 오류를 발생시키지 않기 위해 시계열적 특성을 고려한 전진교차검증을 반드시 적용하여야 한다고 할 수 있다(Snijders,1988).

k-fold 교차검증 vs. 전진교차검증

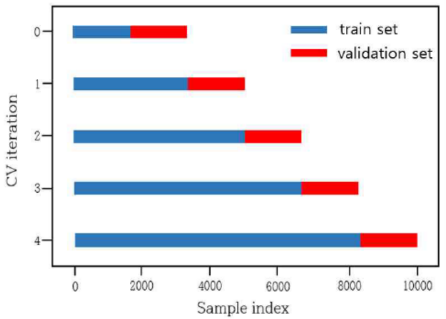
- 비슷한 성능: 대부분의 기계학습 모형에서 소수범주(부도 회사)를 부도로 예측하는 성능이 비슷함.
- 특정 모형에서의 우수성: ROS를 적용한 비선형 SVM과 Random Forest, SMOTE를 적용한 선형 SVM에서 k-fold 교차검증이 상대적으로 우수한 성능을 보임.

문제점 및 해결책

- 논리적 모순: k-fold 교차검증은 미래 데이터를 사용하여 모형을 추정하고 과거 데이터를 통해 부도를 예측하는 모순 발생.
- 이로 인해 모형의 성능이 과대평가될 수 있음

평가지표 및 성능

oversampling 기법 적용 시점에 따른 모형 성능 비교



		oversampling for train and validation set			oversampling for train set only		
		AUC for CV	AUC	Sensitivity	AUC for CV	AUC	Sensitivity
Logistic Regression	ROS	0.7849	0.8624	0.7962	0.7617	0.8624	0.7962
	SMOTE	0.8340	0.8426	0.7488	0.7665	0.8426	0.7488
Linear SVM	ROS	0.8583	0.8693	0.8057	0.7737	0.8688	0.8057
	SMOTE	0.9073	0.8624	0.7867	0.7787	0.8601	0.7820
Non-Linear SVM	ROS	0.8900	0.8337	0.7678	0.8180	0.8337	0.7678
	SMOTE	0.9198	0.8283	0.7204	0.8055	0.8345	0.7678
Random Forest	ROS	0.9318	0.7662	0.5545	0.8481	0.8255	0.7393
	SMOTE	0.9557	0.8376	0.7346	0.8421	0.8438	0.7820
XGBoost	ROS	0.9167	0.8497	0.7820	0.8665	0.8473	0.7867
	SMOTE	0.9384	0.8300	0.7346	0.8525	0.8402	0.7725

* CV = cross validation

□□□ □□□□□□ □□□□ □□□□ □□□□□ □□□□ □□□□□□ □□□□□ □□□□□ oversampling □□□ □□ 검증 내부에서 적용되어야 한다 . 즉, oversampling 기법은 학습세트와 검증세트 전체가 아닌 학습세트에만 적용한 후 □□□□□□ □□□□□□ □□□□□□ □□. □□□□ □□□□ □□ □□□□ □□□□□□ □□ □□□□□□ AUC □□ 시험데이터에서의 AUC 값 즉, 최종 모형의 성능을 대표할 수 있다는 것을 확인할 수 있음

모형성능 비교

- 학습세트와 검증세트 전체에 oversampling 기법을 적용한 기계학습 모형은 oversampling 기법의 종류에 상관없이 교차검증을 통해 얻어진 AUC값에 과적합의 현상이 발생한 것으로 이해할 수 있다.
- 따라서 불균형 데이터에서 과적합 현상을 방지하고 올바른 교차검증을 실시하기 위해서는 oversampling 기법이 교차검증 내부에서 적용되어야 한다.

문제점 및 해결책

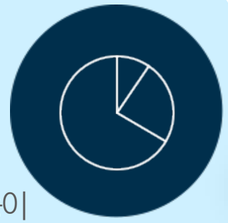
- 즉, oversampling 기법은 □□□□□ □□□□ 전체가 아닌 학습세트에만 적용한 후 □□□□□ □□□□□ □□□□ □□.
- □□□ □□□ □□ □□□□ □□□□□□ □□ □□□□□ AUC □□ □□□□□□□□□ AUC 값 즉, 최종 모형의 성능을 대표할 수 □□□ 것을 확인할 수 있다.

논문 결론

01. 데이터 불균형 문제 해결

- **Oversampling 기법의 필요성**
  원데이터를 그대로 사용할 경우, 정확도가 높아 보이지만 이는 대부분 정상기업으로 분류되기 때문임
- **AUC와 민감도 개선**
  Oversampling 기법 적용 시, 소수범주(부도 기업)의 예측 성능이 상대적으로 개선됨

02. 시계열적 특성 반영



- **전진교차검증의 필요성**
  k-fold 교차검증 사용 시, 미래 데이터를 이용해 과거 데이터를 예측하는 논리적 모순이 발생하여 모형 성능이 과대평가될 수 있음.
- **전진교차검증**
  시계열적 특성을 반영하여 더 신뢰성 있는 모형 구축이 가능함

03. Oversampling 기법 적용 시점

- **학습세트에만 적용**
  학습세트와 검증세트 전체에 적용할 경우, 과적합 문제가 발생함.
- **과적합 방지**
  학습세트에만 oversampling 기법을 적용하고 검증세트는 원데이터 그대로 사용하는 것이 바람직함.

04. 향후 연구 방향

- **부도 데이터 특성 반영**
  부도 데이터의 불균형과 시계열적 특성을 반영한 합리적 모형 추정 과정 확인
- **다양한 oversampling 기법 연구**
  ADASYN, GAN 등 다양한 기법을 적용하여 모형 성능을 개선할 방안 모색
- **비재무적 특성 반영**
  재무적 특성뿐 아니라 시장리스크와 같은 비재무적 특성을 반영하여 예측 성능을 개선할 필요 있음.

논문 재무변수 선별을 위한 t-test 결과

부도예측모형을 위한 설명변수 선정

변수명	설명
asset	총자산
liability	총부채
capital	총자본
cash	현금 및 현금성자산
retained_earnings	이익잉여금
current_asset	유동자산
current_liabilities	유동부채
working_capital	순운전자본
earnings	당기순이익
profit_before_tax	세전계속사업이익
non_current_asset	비유동자산
accounts_receivable	매출채권
sales	매출액
purchase_liabilities	매입채무
operation_fixed_asset	영업고정자산

<표 3> 파생 재무변수

변수명	설명	출처
EQTA	총자본 / 총자산	박종원, 안성만 (2014)
TLTA	총부채 / 총자산	Ohlson (1980)
CLTA	유동부채 / 총자산	Zmijewski (1984)
CLTL	유동부채 / 총부채	김성태, 강충오, 이필상 (2010)
TLEQ	총부채 / 총자본	박종원, 안성만 (2014)
CLCA	유동부채 / 유동자산	Ohlson (1980)
CATA	유동자산 / 총자산	박종원, 안성만 (2014)
FATA	비유동자산 / 총자산	남재우, 이회경, 김동석 (2000)
TRTP	매출채권 / 매입채무	전현우, 정용화, 신동휴 (2011)
CACL	유동자산 / 유동부채	Zmijewski (1984)
lnTA	log(총자산)	Altman et al. (1977)
RETA	이익잉여금 / 총자산	Altman (1968)
NITA	당기순이익 / 총자산	Ohlson (1980)
NISL	당기순이익 / 매출액	남재우 외 (2000)
EBTA	세전계속사업이익 / 총자산	Altman (1968)
CASHTA	현금및현금성자산 / 총자산	Nam et al. (2009)
WCTA	순운전자본 / 총자산	Altman (1968)
SLTA	매출액 / 총자산	Altman (1968)
SLTP	매출액 / 매입채무	김종만, 홍성희 (1999)
SLEQ	매출액 / 총자본	박종원, 안성만 (2014)
SLFA	매출액 / 영업고정자산	정완호 외 (2006)

부도예측모형을 위한 t-test 이후 변수 선택

<표 4> 최종 설명변수

종류	시점	변수
재무변수	0	CLTL, lnTA
	1	WCTA, SLTA
	2	NITA, CASHTA
	3	EQTA, TLTA, CLTA, FATA
	4	TLEQ, CLCA, CACL, RETA, EBTA
증감율 변수	0	총자산, 총부채, 총자본, 당기순이익, 매입채무, EQTA, NITA
	1	세전계속사업이익, 매출액, EBTA
	2	유동부채, 비유동자산, CATA, FATA, WCTA
	3	영업고정자산, CLTA, CLTL, CLCA, CACL, lnTA
	4	현금 및 현금성자산, TLTA, CASHTA, SLTA, SLEQ, SLFA

논문 재무변수 선별을 위한 t-test 결과

재무변수 시차 선택을 위한 변수별 t-test 결과

변수	시차	t	p-value	변수	시차	t	p-value
WCTA	0	2.5613	0.0111	lnTA	0	6.6461	0.0000
	1	3.9727	0.0001		1	7.6737	0.0000
	2	5.1044	0.0000		2	8.8518	0.0000
	3	5.7055	0.0000		3	9.8003	0.0000
	4	4.2190	0.0000		4	9.8561	0.0000
TRTP	0	-1.4022	0.1609	FATA	0	-3.0707	0.0024
	1	-1.4140	0.1574		1	-4.2487	0.0000
	2	-1.4174	0.1564		2	-1.2754	0.2022
	3	-1.4523	0.1464		3	1.9950	0.0473
	4	-1.5872	0.1125		4	2.2948	0.0227
TLTA	0	-11.0132	0.0000	EQTA	0	11.2364	0.0000
	1	-11.6271	0.0000		1	11.6464	0.0000
	2	-11.0145	0.0000		2	10.9153	0.0000
	3	-10.3853	0.0000		3	10.3116	0.0000
	4	-6.2992	0.0000		4	6.2599	0.0000
TLEQ	0	-2.3273	0.0204	EBTA	0	6.8451	0.0000
	1	-0.1664	0.8680		1	3.8679	0.0001
	2	0.5189	0.6044		2	6.5805	0.0000
	3	0.0472	0.9624		3	4.9750	0.0000
	4	0.6114	0.5409		4	2.1719	0.0310
SLTP	0	-1.0170	0.3091	CLTL	0	-0.5088	0.6109
	1	-1.0171	0.3091		1	-1.5362	0.1245
	2	-1.0172	0.3091		2	-1.8888	0.0603
	3	-1.0188	0.3083		3	-2.2558	0.0251
	4	-1.0247	0.3055		4	-2.4059	0.0170
SLTA	0	3.7351	0.0002	CLTA	0	-9.5043	0.0000
	1	5.0793	0.0000		1	-10.1511	0.0000
	2	5.8601	0.0000		2	-9.8589	0.0000
	3	6.0088	0.0000		3	-8.9782	0.0000
	4	4.0688	0.0001		4	-5.1629	0.0000
SLFA	0	-0.2476	0.8046	CLCA	0	-5.7618	0.0000
	1	-0.1958	0.8450		1	-5.6700	0.0000
	2	-1.0089	0.3142		2	-3.9516	0.0001
	3	-0.8552	0.3934		3	-1.5055	0.1337
	4	-0.8308	0.4070		4	1.1451	0.2535
SLEQ	0	−1.1481	0.2509	CATA	0	−0.4962	0.6198
	1	−0.2388	0.8115		1	0.2849	0.7757
	2	0.9250	0.3560		2	1.1915	0.2348
	3	0.4192	0.6755		3	1.8698	0.0629
	4	0.3157	0.7526		4	2.4103	0.0168
RETA	0	10.7194	0.0000	CASHTA	0	2.5652	0.0103
	1	10.6500	0.0000		1	−0.5905	0.5555
	2	8.9711	0.0000		2	6.1769	0.0000
	3	9.4825	0.0000		3	4.3087	0.0000
	4	4.5366	0.0000		4	3.6786	0.0003
NITA	0	6.5415	0.0000	CACL	0	9.6440	0.0000
	1	3.8540	0.0002		1	9.3287	0.0000
	2	6.5933	0.0000		2	8.1422	0.0000
	3	5.0674	0.0000		3	7.9488	0.0000
	4	2.2028	0.0287		4	2.9835	0.0032
NISL	0	−0.9744	0.3310				
	1	1.1559	0.2477				
	2	1.3701	0.1717				
	3	1.1566	0.2474				
	4	1.0055	0.3147				

변수	시차	t	p-value	변수	시차	t	p-value
d_working_capital	0	1.2602	0.2090	d_lnTA	0	7.2059	0.0000
	1	2.4775	0.0140		1	9.1333	0.0000
	2	2.9300	0.0037		2	8.5386	0.0000
	3	1.7544	0.0808		3	6.9127	0.0000
	4	2.5278	0.0121		4	7.0302	0.0000
d_WCTA	0	1.2636	0.2078	d_liability	0	-1.0264	0.3058
	1	2.3771	0.0183		1	1.6204	0.1066
	2	2.9003	0.0041		2	2.0871	0.0381
	3	1.6540	0.0996		3	4.4896	0.0000
	4	2.7642	0.0061		4	5.4331	0.0000
d_TRTP	0	1.3804	0.1675	d_FATA	0	-1.4562	0.1466
	1	1.3804	0.1675		1	-2.3928	0.0175
	2	1.3804	0.1675		2	3.8628	0.0001
	3	1.3804	0.1675		3	3.5766	0.0004
	4	1.3804	0.1675		4	2.4330	0.0158
d_TLTA	0	-6.9845	0.0000	d_EQTA	0	3.0976	0.0022
	1	-6.8620	0.0000		1	2.1867	0.0298
	2	-6.8330	0.0000		2	1.9692	0.0503
	3	-5.2392	0.0000		3	3.2003	0.0016
	4	-3.9099	0.0001		4	4.0673	0.0001
d_TLEQ	0	-1.9179	0.0564	d_EBTA	0	1.8241	0.0681
	1	0.6114	0.5416		1	1.8297	0.0673
	2	-1.2330	0.2189		2	1.8273	0.0677
	3	0.4060	0.6852		3	2.0406	0.0413
	4	-0.4002	0.6890		4	1.7944	0.0728
d_SLTP	0	1.0001	0.3173	d_earnings	0	2.0833	0.0372
	1	1.0001	0.3173		1	2.2579	0.0240
	2	1.0001	0.3173		2	2.2147	0.0268
	3	1.0001	0.3173		3	2.1802	0.0292
	4	1.0001	0.3173		4	2.2603	0.0238
d_SLTA	0	5.1578	0.0000	d_current_liabilities	0	-1.2076	0.2283
	1	-0.9754	0.3305		1	-0.0983	0.9218
	2	4.4725	0.0000		2	1.9212	0.0560
	3	2.4610	0.0139		3	3.2350	0.0014
	4	2.4698	0.0135		4	3.3724	0.0009
d_SLFA	0	5.0564	0.0000	d_current_asset	0	5.6808	0.0000
	1	-0.9846	0.3260		1	8.2205	0.0000
	2	3.6913	0.0002		2	7.9736	0.0000
	3	2.9713	0.0030		3	3.9321	0.0001
	4	2.9831	0.0029		4	2.5243	0.0123
d_SLEQ	0	5.1210	0.0000	d_CLTL	0	0.9804	0.3269
	1	-0.9756	0.3304		1	-2.0086	0.0458
	2	4.5633	0.0000		2	-0.2165	0.8288
	3	-0.9186	0.3583		3	-0.2113	0.8328

	4	-0.9176	0.3588		4	0.4972	0.6196
d_sales	0	4.6950	0.0000	d_CLTA	0	-0.0836	0.9334
	1	-0.7937	0.4283		1	-6.5572	0.0000
	2	-0.7233	0.4695		2	-6.3479	0.0000
	3	4.8716	0.0000		3	-4.4177	0.0000
	4	-0.7693	0.4426		4	-3.3985	0.0008
d_retained_earnings	0	-1.2211	0.2220	d_CLCA	0	0.4983	0.6183
	1	-1.2260	0.2202		1	-6.8460	0.0000
	2	-1.2178	0.2233		2	-5.0385	0.0000
	3	-1.2188	0.2229		3	-1.8360	0.0678
	4	-1.2261	0.2202		4	1.4691	0.1433
d_RETA	0	-1.2118	0.2256	d_CATA	0	1.1036	0.2710
	1	-1.2161	0.2239		1	3.4586	0.0005
	2	-1.2077	0.2272		2	3.6548	0.0003
	3	-1.2095	0.2265		3	3.0160	0.0029
	4	-1.2162	0.2239		4	1.2307	0.2198
d_purchase_liabilities	0	1.8737	0.0610	d_CASHTA	0	3.2154	0.0013
	1	1.9210	0.0547		1	3.2152	0.0013
	2	1.9235	0.0544		2	3.2138	0.0013
	3	2.0340	0.0420		3	3.2137	0.0013
	4	2.0340	0.0420		4	3.2142	0.0013
d_profit_before_tax	0	1.8814	0.0599	d_cash	0	3.2765	0.0011
	1	1.8866	0.0592		1	3.2763	0.0011
	2	1.8843	0.0595		2	3.2750	0.0011
	3	2.0819	0.0374		3	3.2748	0.0011
	4	1.8483	0.0646		4	3.2754	0.0011
d_operation_fixed_asset	0	6.3085	0.0000	d_capital	0	3.3082	0.0011
	1	7.6905	0.0000		1	2.2777	0.0237
	2	7.6382	0.0000		2	1.8062	0.0723
	3	1.7294	0.0852		3	2.4836	0.0138
	4	3.5975	0.0004		4	3.8144	0.0002
d_non_current_asset	0	0.0544	0.9566	d_CACL	0	5.6275	0.0000
	1	-0.2397	0.8108		1	6.8572	0.0000
	2	8.3811	0.0000		2	6.5096	0.0000
	3	5.5259	0.0000		3	3.0519	0.0026
	4	2.6900	0.0077		4	1.9552	0.0519
d_NITA	0	1.8904	0.0587	d_asset	0	6.5066	0.0000
	1	2.1264	0.0335		1	9.1072	0.0000
	2	2.0887	0.0367		2	8.5630	0.0000
	3	2.0575	0.0396		3	5.0955	0.0000
	4	2.1373	0.0326		4	9.2227	0.0000
d_NISL	0	-0.9998	0.3185	d_accounts_receivable	0	-0.3782	0.7056
	1	1.0192	0.3081		1	1.4313	0.1524
	2	1.0020	0.3175		2	1.3990	0.1618
	3	0.9530	0.3417		3	1.3995	0.1617
	4	-0.5306	0.5957		4	1.3929	0.1637

05 마무리하며... 느낀점

01. 소감

관심분야인 '□□□□□□□□□□□ 선행연구 Case STUDY 를 하면서

□□□ □□□□ □□□□ □□□(Neural Network), □□□□□(DNN) □□□ □□ □□□ □□

이번 과제는 본 학기 수업내용에 맞추어 본 논문을 찾아 CASE STUDY 부분에 제약 사항이 있었음

논문 분석을 통해 그동안 배운 내용을 논문의 내용과 비교하면 다시 복습하고 체득하는 계기였음

향후 다양한 신경망 활용 사례 분석을 위한

'□□□ □□□' □□ !!

~

Aa 논문명	모델	모델 요약	성능	링크	년도	문제유형
머신러닝과 인공신경망을 활용한 수출제조기업 신용등급 예측연구	로지스틱 회귀분석, XGBoost, ANN	통계기반 변수선택 vs 트리기반 변수선택 후 모델링 → 성능비교	ANN(ACC 0.839) > XGBoost(0.790) > RF(0.779) > LR(0.758)	dbpia.co.kr/jou...426394	2023	신용등급
머신러닝 기반 기업부도 위험 예측모델 검증 및 정책적 제언: 스태킹 앙상블 모델을 통한 개선을 중심으로	• 랜덤 포레스트 • 다층 퍼셉트론 (Multiple Layers Perceptron) • 합성곱 신경망 • 스태킹 앙상블			dbpia.co.kr/jou...365028	2020	부도예측

심층신경망의 설명가능성과 하이퍼파라미터 특성에 관한 연구 : 중소기업 신용평가를 중심으로	DNN, LR		AUC: DNN > LR	dbpia.co.kr/jou...275175	2022	부도예측
기계학습 기반 기업신용정보 분석을 통한 채무불이[행] 예측	• 모수적 - LR, LDA • 비모수적 - KNN, NB / 트리계열 방법론 - DT, RF, XGBoost, ExtraTreeClassifier	DNN - 16개 layer, 출력층 제외 각 층의 노드는 256개, 출력층에서 0.5 드롭아웃	전체 세그먼트 - DNN > DT 개인 세그먼트 - DNN > ET 법인 세그먼트 - DNN > RF	korfin.org/sub...page=2	2021	부도예측
TabNet을 활용한 딥러닝 성능 비교와 설명가능한 AI 활용성에 대한 연구 : 기업 신용평가 모형을 중심으로	TabNet+복합샘플링	TabNet에 데이터 샘플링 기법인 SMOTE+ENN 추가	LR, MLP대비 높은 성능(AUC 0.868, F1 0.88, FNR 0.1919)	dcollection.sogang.ac.kr/dco...Param=	2022	부도예측
다양한 앙상블 모델을 이용한 개인신용평가 모델 개발	DNN, Lasso logistic, Gradient Boosting, Model1, Model2	Model1: 모델을 다른 샘플링 된 데이터 이용해 학습 후 확률의 평균을 최종 확률로 사용	Model1(0.76) > Gradient Boosting(0.75) > Lasso logistic(0.72) > Model2(0.71) > DNN(0.69)	riss.kr/sea...%8D%B8	2020	신용등급

02. 소감

모델 중심의 인공지능 개발 VS 도메인 중심의 인공지능 개발

머신러닝 논문을 검색하고 읽어 보는 과정에서 생각보다 머신러닝 개발 과정에서 해당 분야의 데이터와 해당 도메인의 기술과 지식이 부족하여 편향된 분석을 하는 것을 볼 수 있었음.
인공지능의 모델 구조와 요소 기술의 발전도 중요하지만 해당분야의 높은 이해도를 기반으로 데이터와 Expert Knowleadge를 활용한 솔루션(모델)을 개발하는 것도 중요할 것 같습니다.

‘데이터와 도메인지식의 중요성’

모델 중심 개발과 도메인(데이터) 중심 개발의 성능 개선 차이

	Steel Defect Detection	Solar Panel	Surface Inspection
Baseline	76.2%	75.68%	85.05%
Model-Centric	+0% (76.2%)	+0.04% (75.72%)	+0.00% (85.05%)
Data-Centric	+16.9 (93.1%)	+3.06% (78.74%)	+0.4% (85.45%)

출처 | Youtube, A Chat with Andrew on MLOps: From Model-centric to Data-centric AI

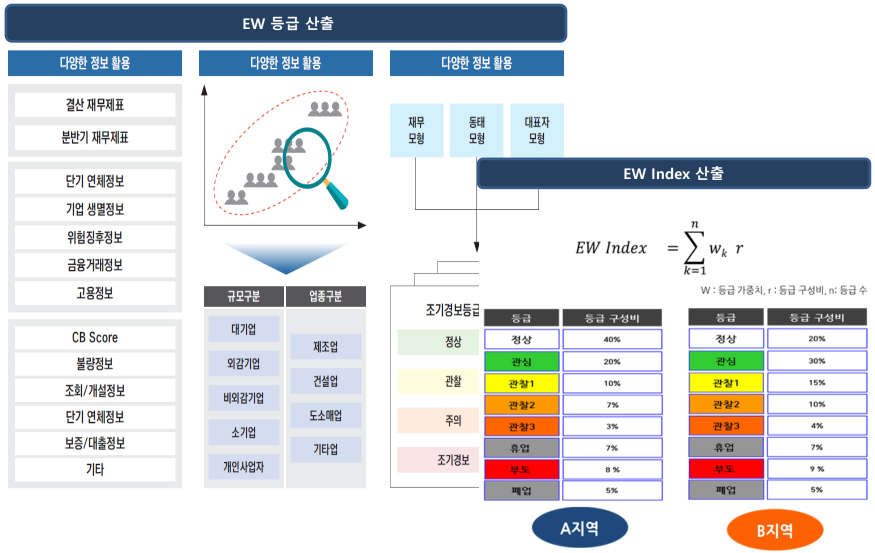
AI 시스템 = 모델 + 데이터 + 휴리스틱
도메인 중심 AI 개발

	Model - Centric	Domain – Centric
Training	Model Build & Evaluation	Pre-Trained Model (+ Model Build & Evaluation)
Data-set	Assigned	User(Engineer) Define
Label		
Objective	Assigned (Classification, Regression etc.)	
Result	Transformer, BERT, GPT, LSTM, Resnet, Alexnet, VGG	

CASE STUDY(실무 사례)

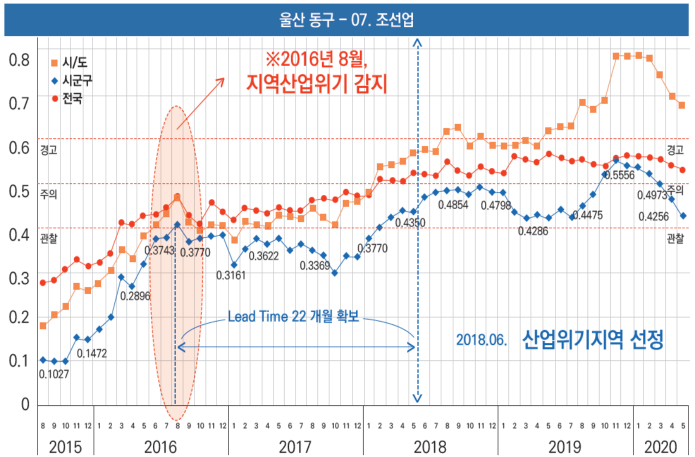
- EW(□□□□) □□□ □□□□, □□□□ □□, □□□ □□, □□□□, □□□/□□□ □□, □□□□ □□ □□□□, □□, □□, □□, □□, □□ □□□□ □□
- EW Index는 등급별 가중치를 적용하여 □□□ □□□ 계산되며 , 추정 부도율을 의미 □□ □□ □□□ □□□□ □□□□□ □□□ □□□

- 울산 동구 지역의 조선업은 2018년 6월 산업위기지역으로 선정되었으며 , EW Index는 2016년 8월부터 위험수준의 임계값을 상회하는 것을 보여주고 있어 EW Index의 □□ □□ □□ □□□ □□



산업위기 특별대응지역
- 산업 구조 조정 등 경제위기로 지역 내 대규모 휴폐업/실직 등 위기에 봉착한 경우, 범부처가 합동으로 지원할 수 있는 경제 산업분야의 특별 재난 지역

- 거제, 통영/고성, 영암/목포/해남, 울산 동구는 고시 제4조 제1항 및 2항에 따른 지정 기준을 충족하여 현장실사 결과 지역 경제 침체가 확인되어 지정 추진
- 창원 진해구는 STX조선 구조조정 등으로 고용이 악화되었으며('17. 4. 5. 고용위기지역 지정) 현장실사 결과 지역 경제 침체가 확인되므로 고시 제4조 5항에 따라 지정 추진



EW 등급(조기경보) 데이터를 활용
부실기업 분류 및 미래 부실 기업 라벨링

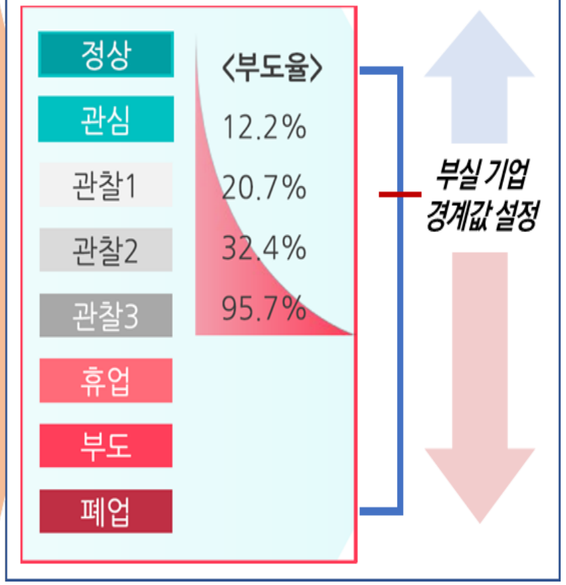
예측분석 마트

EXAMPLE

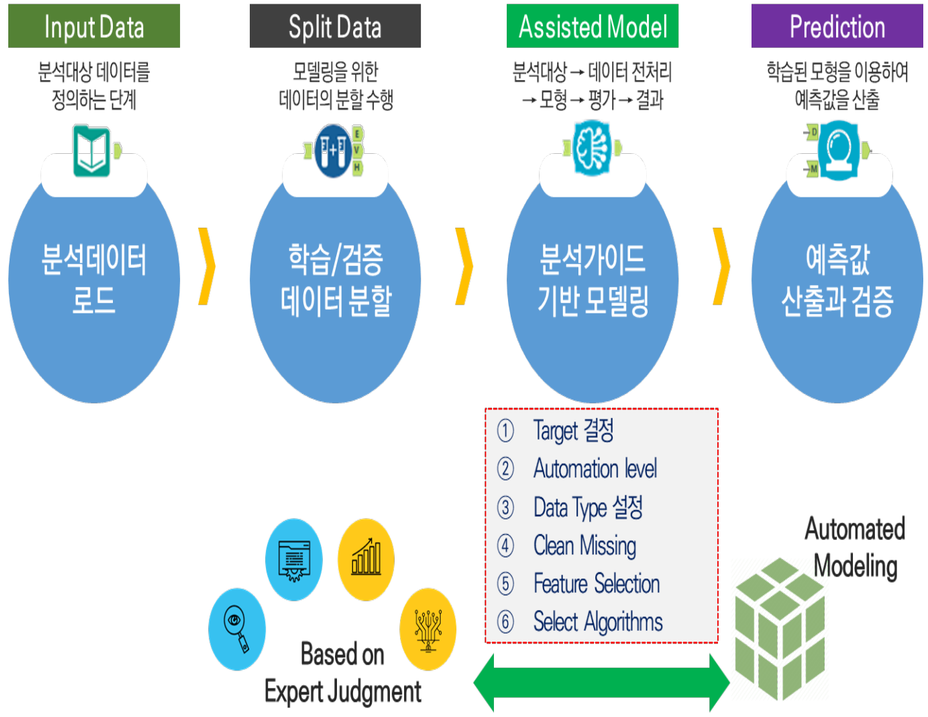
업체명	기준연도	기업공개	기업규모	매출액	매출액 증감률
A	2020	외감	중기업	67363973	-0.2%
A	2021	외감	중기업	57353400	-5.2%
B	2020	일반법인	소상공인	1744853	2%
B	2021	일반법인	소상공인	2754541	10%
C	2020	유가증권시장	한시성중소기업	14066103	5.4%
C	2021	유가증권시장	한시성중소기업	10060105	-21.4%
D	2020	외감	중기업	23413499	22.4%
D	2021	외감	중기업	21410400	-1.4%

LABELING

미래 EW 등급 현황



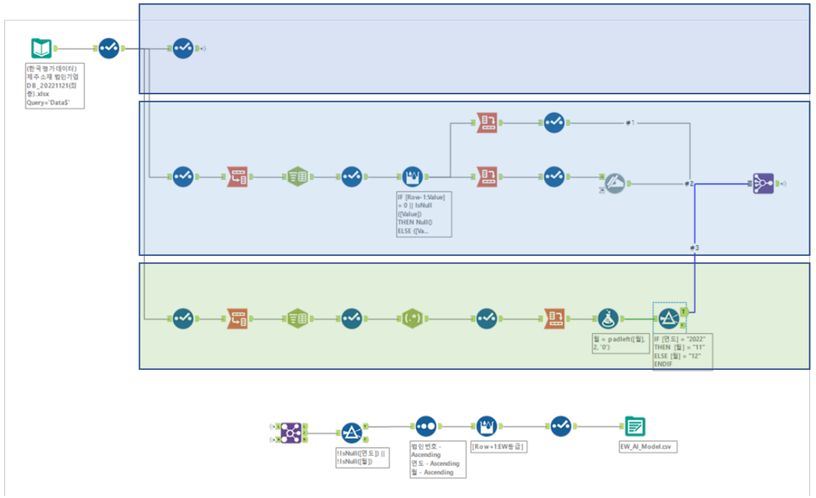
잠재부실 기업예측 Auto ML활용 돌려보기



06 CASE STUDY (유사 분석 중심 파일럿 테스트사례)

원천데이터 가공 및 라벨링 WorkFlow

01.



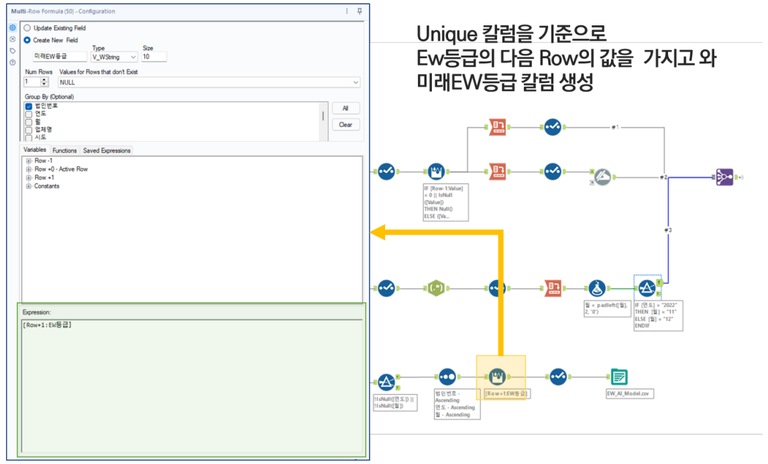
기업 속성 및 행위 정보

기업 연도별 시계열 정보

기업 월별 시계열 정보

라벨(미래 EW 등급) 생성

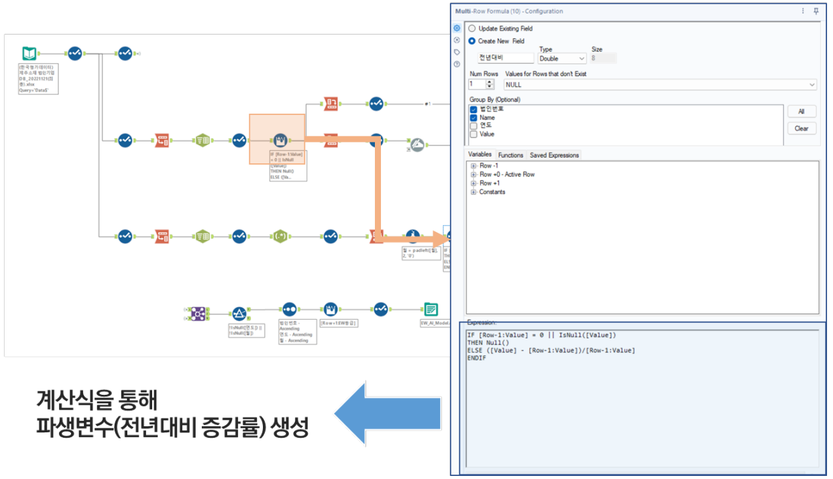
03.



Unique 칼럼을 기준으로
Ew등급의 다음 Row의 값을 가지고 와
미래EW등급 칼럼 생성

파생 변수 생성

02.



계산식을 통해
파생변수(전년대비 증감률) 생성

잠재 부실 기업 예측 – 모델링

04.

EXAMPLE

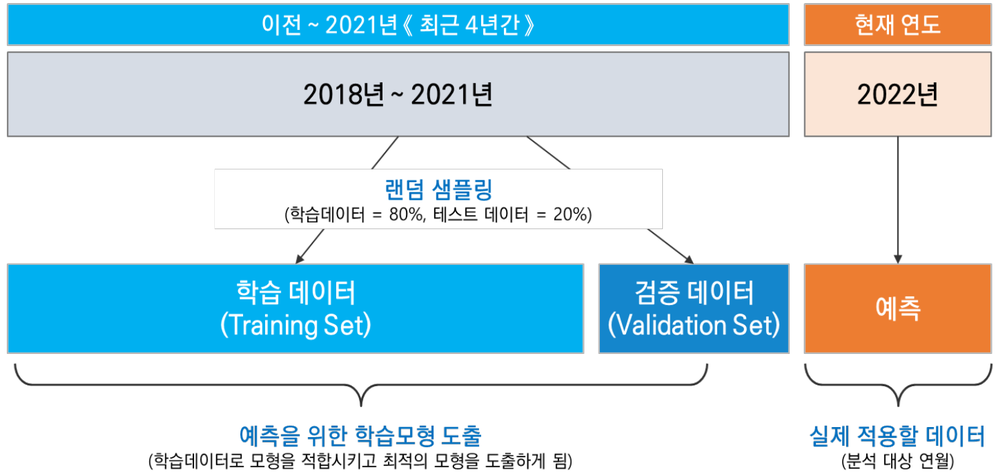
Label

업체명	기준연도	기업공개	매출액	매출액 증감률	현재 EW 등급	미래 EW 등급
A	2020	외감	67363973	-0.2%	1	1
A	2021	외감	57353400	-5.2%	1	0
B	2020	일반법인	1744853	2%	0	0
B	2021	일반법인	2754541	10%	0	1
C	2020	유가증권시장	14066103	5.4%	1	0
C	2021	유가증권시장	10060105	-21.4%	1	1
D	2020	외감	23413499	22.4%	1	0
D	2021	외감	21410400	-1.4%	0	1

1. 데이터 샘플링 — 학습용 데이터와 평가용 데이터를 분리
2. 데이터 정규화 — 데이터의 단위를 동일화(Scale) 하기 위해 데이터 정규화
3. 결측값 대체 — 수집하기 어려운 항목, 비어 있는 값들을 분석에 활용하는 기준
4. 데이터 변환 — 결과해석의 용이성, 모형의 성능을 고려한 데이터 변환
5. 변수 선택 — Target과의 통계적 유의성을 분석하여 분석활용 피처를 선별
6. Modeling — 예측을 위한 다양한 모형들을 적용하여 최적 모형 도출

06 CASE STUDY (유사 분석 중심 파일럿 테스트사례)

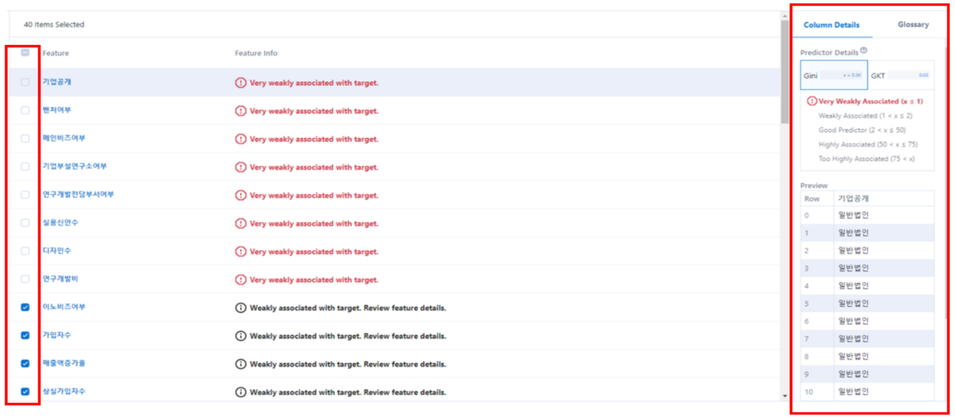
05. 모델링 데이터 샘플링



06. 변수 선택

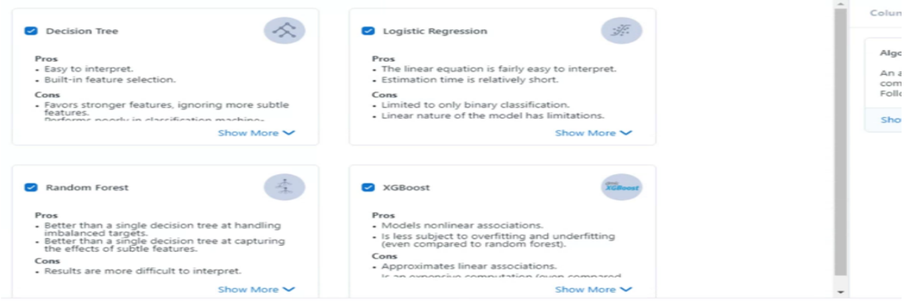
- 컬럼 별 지니 불순도(Gini)와 Goodman-Kruskal Tau(GKT)가 계산되어 Target과의 연관성 측정
- 2가지 지표 중 높은 값을 기준으로 규칙 기반으로 Target 과의 연관성 분류
- 지표가 매우 높거나 낮은 컬럼에 대해서는 자동으로 예측모델에서 사용되지 않도록 설정

Filter 기반의 통계량 산출후 제외 여부 추천
→ "자동추천+선택"



07. Modeling

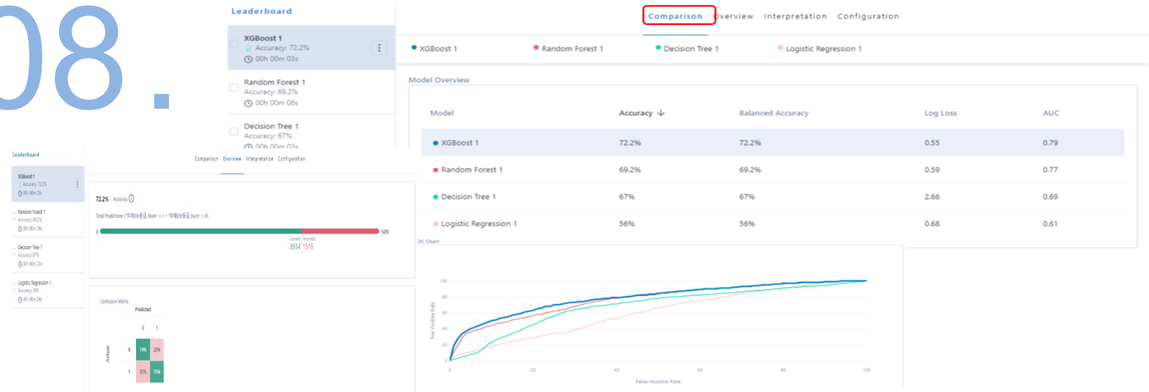
- 사용할 알고리즘을 선택하고 AutoML 시작
- 앞서 선택한 머신러닝 작업에 따라서 알고리즘 선택
  1) Classification : Decision Tree, Logistic Regression, Random Forest, XGBoost
  2) Regression : Decision Tree, Linear Regression, Random Forest



08. Modeling

예측성능지표 기반의 최적 모형 선정
- 선택한 알고리즘 별 학습 시간 및 각종 예측 성능 지표 비교

예측성능지표를 예측모형별로 비교 검토
→ "AUC, Log Loss 등"



06 CASE STUDY (유사 분석 중심 파일럿 테스트사례)

테스트 데이터 예측 및 검증(1/2)

- 기본 모형은 2018년~2020년으로 학습후 2021년 예측
- 시나리오 기반으로 (1)추가 인자 반영, (2)다양한 학습 데이터 기준에 따라 최적 모형 선정을 진행

시나리오 구분	대상 기업	학습 셋	테스트 셋	특이사항	Model*	Metric				
						Accuracy	Precision	Recall	F1-score	AUC
S1	연도별 미래 Ew 등급 존재 기업	2018~2020	2021	전체값	LR	0.856274	0.170732	0.008878	0.016878	0.577279
					DT	0.86306	0.507201	0.513633	0.510397	0.782392
					RF	0.861033		0		0.817611
					XGB	0.87619	0.551313	0.585923	0.568091	0.82841
S2				언더 샘플링	LR	0.358565	0.166472	0.902346	0.281086	0.597368
					DT	0.766038	0.3324	0.677869	0.446067	0.755748
					RF	0.836711	0.44321	0.682942	0.537559	0.812414
					XGB	0.81098	0.401045	0.729867	0.517652	0.822376
S3		2020		전체값	LR	0.854071	0.193798	0.015853	0.029308	0.581495
					DT	0.858477	0.492072	0.570704	0.528479	0.759706
					RF	0.869492	0.769663	0.086874	0.156125	0.815285
					XGB	0.877864	0.554854	0.612555	0.582278	0.823873
S4				언더 샘플링	LR	0.362619	0.167841	0.906151	0.283223	0.603967
					DT	0.749119	0.317424	0.700063	0.436795	0.744596
					RF	0.83301	0.436042	0.687381	0.533596	0.818864
					XGB	0.811861	0.401134	0.717819	0.514662	0.820465

테스트 데이터 예측 및 검증(2/2)

- 최종 시나리오(AUC를 기준으로 선정된 시나리오)를 통해
- 2018~2021년 각 연도별 데이터를 예측하고 각각의 평가지표 확인

시나리오 구분	대상 기업	학습 셋	테스트 셋	Model*	Metric				
					Accuracy	Precision	Recall	F1-score	AUC
T1	연도별 미래 Ew 등급 존재 기업	2018~2020 전체값	2018	LR	0.586785	0.132225	0.600719	0.216742	0.621507
				DT	0.754878	0.218509	0.611511	0.32197	0.774109
				RF	0.815474	0.276923	0.582734	0.375435	0.82057
				XGB	0.912701	0.870968	0.097122	0.174757	0.835625
T2			2019	LR	0.406571	0.134222	0.796047	0.229712	0.616769
				DT	0.565265	0.178316	0.806828	0.29208	0.713522
				RT	0.474883	0.155674	0.841869	0.262759	0.709726
				XGB	0.616499	0.190325	0.75292	0.303843	0.750269
T3			2020	LR	0.384921	0.155577	0.850075	0.263018	0.626154
				DT	0.802652	0.366629	0.726387	0.487302	0.823025
				RF	0.884437	0.544248	0.645427	0.590535	0.863201
				XGB	0.859466	0.471689	0.736882	0.57519	0.873853
T4			2021	LR	0.856274	0.170732	0.008878	0.016878	0.577279
				DT	0.86306	0.507201	0.513633	0.510397	0.782392
				RT	0.861033		0		0.817611
				XGB	0.87619	0.551313	0.585923	0.568091	0.82841

예측분석 모델 활용 방안

- 잠재부실기업 예측 데이터 및 통합 데이터 셋을 활용, 밀집지역별 잠재부실기업 현황 파악 대시보드 구성
- 업종별 위험등급 변화 현황 및 지표들을 확인하여 지자체 주요 업종에 대한 위험도 파악
- 밀집지역별 잠재부실기업 비중, 매출액 증감률, 고용보험 가입자 수 증감률 등 밀집지역 모니터링을 위한 정보 제공

강사님의 열정적인 강의
감 사 합 니 다