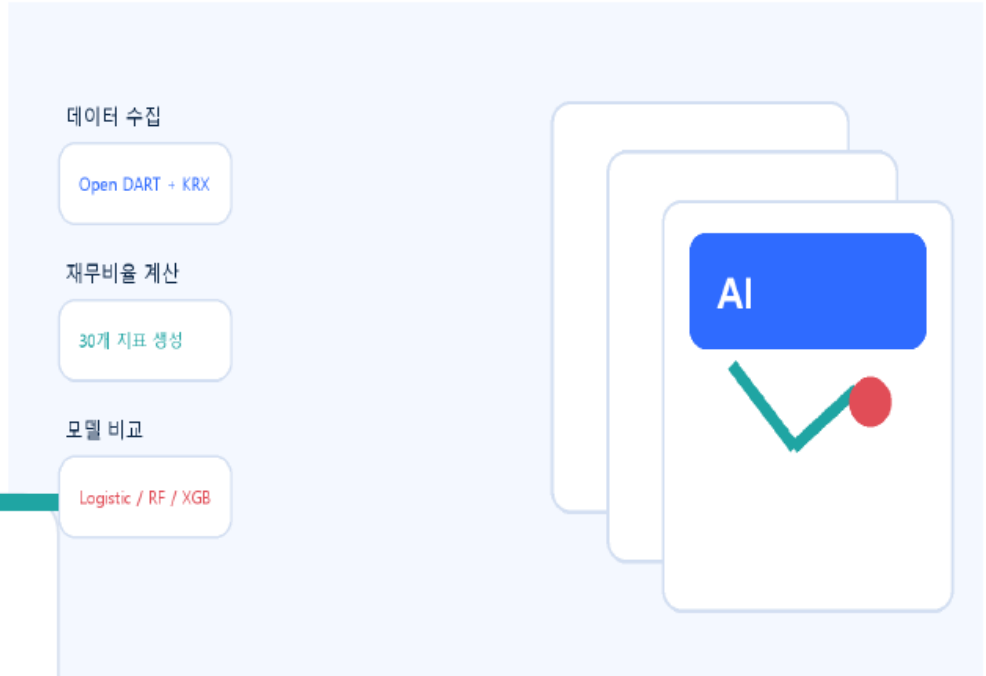
상장폐지 예측 프로젝트

데이터 수집부터 모델 비교와 개선 방향까지

상장기업과 상폐기업 재무비율 기반 AI 분류

- Open DART와 KRX 기반 데이터 수집 구조를 구축했습니다.
- 재무제표에서 30개 재무비율을 계산해 학습용 CSV를 만들었습니다.
- Logistic, Random Forest, XGBoost를 학습해 성능을 비교했습니다.
- 현재 한계와 다음 단계 개선 방향까지 문서화했습니다.



학습 행 수
22,684
정제 후 최종 학습 데이터

상폐기업
273
전체의 1.2% 수준

비교 모델
3개
Logistic / RF / XGB

프로젝트 개요

우리가 만든 전체 흐름

1. 데이터 수집

Open DART와 KRX를 이용해 상장기업과 상폐기업 재무제표를 연도별 JSON으로 수집

→

2. 비율 계산

재무제표 원본에서 30개 재무비율과 보조 금액 컬럼을 계산해 CSV 생성

→

3. 모델 학습

분류 모델 3개를 학습하고 그룹 기반 교차검증으로 성능 비교

→

4. 결과 분석

문제점과 개선 방향을 문서화하고 예측 결과 CSV 저장

- 현재 버전은 데이터 수집, 가공, 학습, 예측이 하나의 흐름으로 연결되어 있습니다.
- 핵심 스크립트는 dart_financial_downloader.py, financial_ratio_calculator.py, delisting_model.py 입니다.
- 이번 실험 목적은 완벽한 실무 모델 완성이 아니라, 상장폐지 예측의 초기 가능성과 한계를 확인하는 것입니다.

데이터 수집

Open DART + KRX 기반 원천 데이터 확보

수집 연도 범위

2015 - 2025

연도별 정기보고서 중심

수집 대상

상장 + 상폐

학습용 양성/음성 동시 확보

- Open DART API에서 재무제표와 기업 메타 정보를 내려받고, KRX 목록으로 상장/상폐 상태를 구분했습니다.
- 연결재무제표(CFS)를 우선 사용하고, 없을 때만 별도재무제표(OFS)를 사용했습니다.
- 결과는 기업별, 연도별 JSON 파일로 저장되어 이후 재무비율 계산의 입력이 됩니다.

- 핵심 스크립트: dart_financial_downloader.py
- 주요 출력: downloads/dart_financials/상장기업, downloads/dart_financials/상폐기업
- 의미: 현재 상장기업뿐 아니라 상폐기업도 같이 모아야 분류 모델 학습이 가능합니다.

재무비율 계산과 데이터셋

재무제표를 모델이 읽기 좋은 형태로 변환

성장성
예시 3개
총자산증가율
매출액증가율
순이익증가율

수익성
예시 3개
매출액순이익률
자기자본순이익률
총자본영업이익률

활동성
예시 3개
매출채권회전율
재고자산회전율
총자본회전율

안정성
예시 3개
부채비율
유동비율
자기자본비율

- financial_ratio_calculator.py에서 원시 재무제표를 읽어 재무비율 30개와 진단용 금액 컬럼 7개를 계산했습니다.
- 최종 학습용 CSV는 45개 컬럼이며, 모델에는 재무비율 30개만 feature로 사용했습니다.
- 예측 테스트용으로 financial_ratios_2024_test.csv도 따로 준비했습니다.

총 컬럼 수
45
메타 8 + 비율 30 + 금액 7

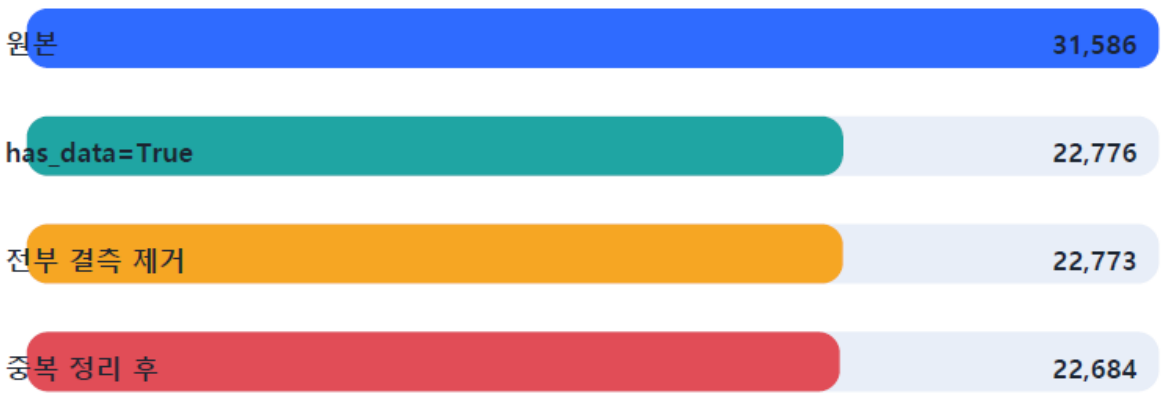
모델 입력
30개
재무비율만 feature 사용

테스트 파일
2024
예측용 CSV 별도 준비

데이터 정제 결과

원본에서 실제 학습 데이터로 가는 과정

정제 단계별 행 수



최종 학습 행 수
22,684
중복 정리와 결측 제거 반영

기업 수
2,690
그룹 분할에 사용

상폐기업
273
양성 샘플

상장기업
22,411
음성 샘플

- has_data=False 행과 30개 비율이 모두 비어 있는 행을 제거했습니다.
- 같은 종목코드와 연도가 중복일 때는 상폐 우선 규칙으로 하나만 남겼습니다.
- 최종적으로 22,684행으로 정리되었고, 이 중 상폐기업은 273행입니다.

핵심 문제 1 - 클래스 불균형

상폐기업이 너무 적어서 모델이 학습하기 어려운 구조

클래스 불균형 시각화

상폐기업 273개 전체의 1.2%

상장기업 22,411개

상폐기업 비율이 너무 작아서 모델이 상장기업 쪽으로 치우치기 쉬운 구조입니다.

현상

상폐기업 비율이 약 1.2%라서 모델이 대부분 상장기업만 보게 됩니다.

→

결과

Random Forest는 상폐기업을 하나도 못 잡았고, XGBoost도 거의 못 잡았습니다.

→

해석

데이터가 적은 양성 패턴을 못 배우거나, 잡으려 하면 오탐이 급증합니다.

모델링 파이프라인

전처리, 모델 3종, 그룹 기반 교차검증

전처리

has_data=True, 전부 결측 제거, 중복 정리, 중앙값 대치

→

모델

Logistic Regression, Random Forest, XGBoost

→

검증

StratifiedGroupKFold 5-fold로 같은 기업이 train/test에 같이 들어가지 않도록 분리

→

산출물

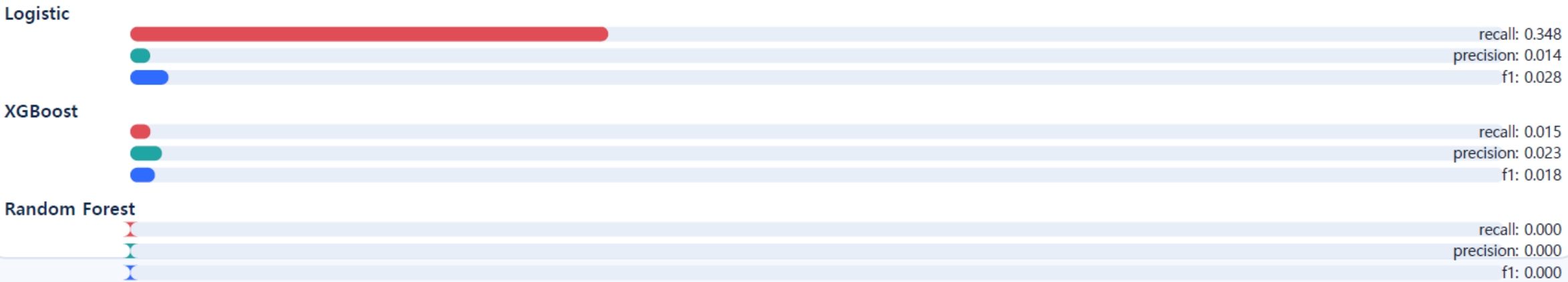
metrics_summary.csv, comparison_report.md, predictions_2024.csv

- 학습 명령: python .₩delisting_model.py train --input .₩downloads₩dart_financials₩financial_ratios_2015_2025.csv --output-dir .₩artifacts₩delisting
- 예측 명령: python .₩delisting_model.py predict --model-dir .₩artifacts₩delisting --input .₩downloads₩dart_financials₩financial_ratios_2024_test.csv --output .₩artifacts₩delisting₩predictions_2024.csv
- 평가 지표는 Recall, Precision, F1, ROC-AUC, PR-AUC를 사용했고, 상폐기업을 놓치지 않는 Recall을 가장 중요하게 봤습니다.

모델 성능 비교

3개 모델 중 Logistic이 상대적으로 가장 나은 결과

모델별 핵심 지표



베스트 모델

Logistic

Recall 기준으로 가장 높음

Logistic Recall

0.348

상폐기업 273개 중 95개 탐지

Logistic Precision

0.014

오탐이 매우 많음

핵심 해석

최고지만 약함

3개 중 최고 != 실전형

결과 해석

왜 현재 모델을 바로 실전에 쓰기 어려운가

Logistic 혼동행렬 해석

	실제 상장	실제 상폐
	TN 15,932	FP 6,479
	FN 178	TP 95

- Logistic은 상폐기업 273개 중 95개를 맞췄지만, 178개를 놓쳤습니다.
- 동시에 상장기업 6,479개를 상폐기업이라고 잘못 경고했습니다.
- 즉, 현재 상태로는 실무 최종 판정용보다 1차 위험 탐지용에 더 가깝습니다.

좋은 점

데이터 수집부터 예측까지 전체 파이프라인을 만들었고, 재무비율만으로 어느 정도 신호가 있다는 점을 확인

→

한계

오탐이 많고 미래 예측 구조가 아니라서 실제 운영 지표와 차이가 큼

→

현재 위치

실전용 완성 모델보다 1차 프로토타입, 방향 확인용 시스템에 가까움

왜 한계가 생겼는가

지금 구조에서 생기는 대표 문제 4가지

문제 1

상폐기업이 너무 적음

양성 데이터가 273개뿐이라 불균형이 심함

문제 2

라벨 정의 한계

현재 상태 분류에 가까워 미래 상폐 예측과 완전히 같지 않음

문제 3

변수 부족

감사의견, 관리종목, 거래정지 같은 핵심 신호가 빠짐

문제 4

오탐 과다

위험 기업을 잡으려 하면 정상 기업 경고가 급증함

- 결국 지금 성능 문제는 모델 알고리즘 하나의 문제가 아니라 데이터 구조와 문제 정의의 문제입니다.
- 따라서 다음 단계에서는 모델만 바꾸기보다 라벨 재설계, 데이터 보강, 보정 기법, 시간 기준 검증을 먼저 개선해야 합니다.

재무제표 데이터 보정

모델 성능을 높이기 위해 적용할 수 있는 전처리

결측치 보정

중앙값 또는 업종/연도별 중앙값 대치 + 결측 플래그

→

이상치 완화

Winsorization, clipping, 금액 컬럼 log 변환

→

파생변수

전년 대비 변화, 2년 연속 적자, 부채 급등 여부

→

불균형 대응

class_weight, scale_pos_weight, threshold tuning

- 재무제표 데이터라고 해서 원본을 그대로 쓰는 것이 일반적인 것은 아닙니다. 모델용 데이터셋은 보정을 거치는 경우가 훨씬 많습니다.
- 다만 전체 데이터로 기준을 잡으면 데이터 누수가 생기므로, 중앙값과 이상치 기준은 반드시 train 데이터에서만 계산해야 합니다.
- 현재 프로젝트에서는 결측치 보정, 이상치 완화, 시계열 파생변수, 불균형 대응이 우선순위가 높습니다.

개선 로드맵 - 라벨 재설계 우선

다음 연도 상폐 여부 예측 구조를 먼저 확정

목표 변경

현재 상태 분류가 아니라 다음 연도 상폐 여부 예측으로 문제를 다시 정의

→

라벨 규칙

예: 2021년 재무비율을 입력으로 넣고 2022년 실제 상폐 여부를 정답으로 연결

→

완료 기준

상폐 시점 정의와 라벨 작성 규칙이 문서로 정리되고 한 줄로 설명 가능해야 함

해야 할 일

상폐 시점 정의

실제 상폐 연도를 표로 정리하고
직전 연도를 양성으로 둘지 규칙 확정

예시 구조

2021 -> 2022

2021 재무비율로 2022 상폐 여부를 맞히는
미래 예측형 라벨 구조로 변경

추천

1년 뒤 상폐

가장 먼저 다음 연도 상폐 여부 버전부터
만들어 기준 모델을 다시 학습

- 현재 구조는 기업이 지금 상장/상폐 그룹인지 분류하는 데 가까워서, 실제로 쓰고 싶은 미래 예측 문제와 차이가 있습니다.
- 그래서 먼저 상폐기업의 실제 상폐 연도를 정리하고, 어떤 연도 데이터를 넣었을 때 몇 년 뒤 상폐를 맞히는지 라벨 규칙을 명확히 해야 합니다.
- 이 작업이 끝나야 이후의 추가 변수 결합, 시간 기준 검증, threshold tuning도 의미 있게 진행됩니다.