[기후금융] 재무정보를 이용한 상장폐지 -

✅ **필수 과제**

- 모델 만들기, 최적화, predict_prob
- earning 변수추가
- 산업분류 코드 C1, C2z…

▼ 1. 동일 사업자번호&연도 중복 전처리

- '현진소재', '현진소재 2공장'과 같이 일치하는 사업자번호를 가지는 동일한 회사인 경우 → 같은 연도 배출량 데이터끼리 sum
- 사업자번호를 포함하는 유료 데이터에서 잘못 매칭한 데이터 오류인 경우 → drop 또는 replace

▼ 중복 데이터 목록

- **drop할 기업**
  - 영풍제지(주), 영풍전자(주)
  - 평택시, 페이퍼코리아 세방전지(주)
  - 주식회사 대유글로벌, 주식회사 대창
  - 화영운수(주), 화인베스틸
  - 주식회사 영풍, 주식회사 영흥산업환경 삼성물산 주식회사
  - (주)대우건설, (주)대욱케스트
- **replace할 기업**
  - 포스코에너지(주), 포스코강판(주), 포스코스틸리온(주)
  - KPX 케미칼(주), KPX그린케미칼(주), 그린케미칼 주식회사
  - 롯데쇼핑 주식회사, 주식회사 호텔롯데
- **sum할 기업**
  - 삼정펄프(주)함안공장, 삼정펄프(주)
  - (주)금비, (주)금비 이천공장
  - 현진소재(주) 2공장, 현진소재(주), 현진소재(주) 양산공장, 현진소재주식회사
  - (주)세아제강, (주)세아제강 군산공장
  - 세방전지(주) 창원공장
  - 일진머티리얼즈 주식회사, 일진머티리얼즈 주식회사 조치원공장
  - SK이노베이션(주), SK이노베이션(주) 증평공장, SK이노베이션㈜
  - 신대양제지(주), 신대양제지반월(주)
  - 일신방직(주) , 일신방직(주), 일신방직(주) 광주2공장
  - (주)에코프로비엠, (주)에코프로비엠 1공장

  - 예시

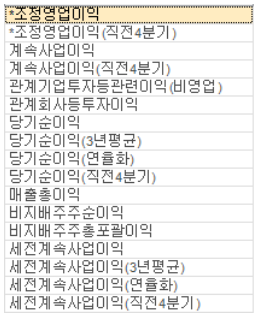
	non_cor_name	year	ctgy	ctgy_detail	grngas	enrg	cor_name	bsn_code	etc
458	현진소재주식회사	2011	사업장	기계	103,742	2,022	현진소재	603-81-06646	
459	현진소재(주) 2공장	2011	사업장	기계	NaN	NaN	현진소재	603-81-06646	
1057	현진소재(주)	2012	사업장	기계	62,278	1,231	현진소재	603-81-06646	
1058	현진소재(주) 2공장	2012	사업장	기계	NaN	NaN	현진소재	603-81-06646	
1626	현진소재(주)	2013	사업장	기계	57709	1140	현진소재	603-81-06646	
1627	현진소재(주) 2공장	2013	사업장	기계	NaN	NaN	현진소재	603-81-06646	
1628	현진소재(주) 양산공장	2013	사업장	기계	NaN	NaN	현진소재	603-81-06646	
2217	현진소재(주)	2014	업체	기계	69288	1370	현진소재	603-81-06646	
2606	현진소재(주)	2015	사업장	기계	59226	1174	현진소재	603-81-06646	
3762	현진소재(주) 2공장	2021	사업장	기계	4,158	83	현진소재	603-81-06646	

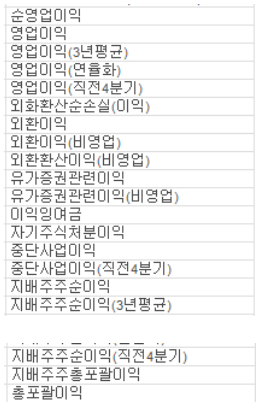
▼ 2. 이익(earning) 변수추가

➕ **추가한 변수 (손익계산서 주요 항목)**

- 매출총이익
- 영업이익
- 이익잉여금

- **FnGuide 이익관련 계정항목**





▼ 3. 응용 모델

> **(데이터 원형(n=0) | 데이터 원형 row로 나열 |변화율 | 지수함수 조합) + (주식 데이터 유무) + (최적 파라미터 n 탐색)**
>
> - train data에서 `['code_label']` 제외
> 각 기업 별로 어차피 하나의 행만 가지므로 의미없는 변수이며, 기업코드가 갖는 범주형 데이터가 서수적 영향을 미치는 것을 제거(과적합 방지)
> - train data에 이익 변수 추가
> 매출총이익`['gross_profit']`, 영업이익`['operating_profit']`, 이익잉여금`['retained_earning']`

- **단독 모델**: 데이터 원형/row로 나열/변화율 데이터 형태/지수함수 데이터 각각으로만 train data를 구성하는 것
- **응용모델**: 데이터 원형, row로 나열, 변화율 데이터 형태, 지수함수 데이터 형태를 조합해 train data를 구성하는 것

▼ [결과 비교] 이익변수 유무에 따른 단독모델 성능

> 📌 **이익변수 추가X vs 이익변수 추가O**
>
> - **주식 데이터O**: 전체적으로 이익변수 추가O 한 모델이 더 성능 Good
> - **주식 데이터X**: 변화율(n=3,6)을 제외하고 이익변수 추가O 한 모델 성능 Good
>
> ⇒ 결론: 이익변수 유무에 따른 드라마틱한 성능 변화는 없으며, 이익변수는 후에 기후정책 변화 충격을 주는 단계에서 사용해야 하는 변수이므로 **이익변수 추가를 default**

- **이익변수 추가X 단독모델(이전 결과 보고 때의 정확도)**

<단독 모델> * 주식 데이터O	데이터 원형	데이터 원형 (n=2)	변화율 (n=5)	지수함수 (n=3)
정확도	0.9392	0.9191	0.9097	0.9146
상장폐지를 상장유지로 예측 (오답률 %)	27 / 506 (5%)	(4.3%)	(4.2%)	(4.5%)
상장유지를 상장폐지로 예측 (오답률 %)	**10 / 103 (9%)**	**(27.8%)**	(33.0%)	(30.1%)

<단독 모델> * 주식 데이터X	데이터 원형	데이터 원형(n=2)	변화율 (n=3, 6)	지수함수 (n=2)
정확도	0.938	0.9109	0.9310	0.9442
상장폐지를 상장유지로 예측 (오답률 %)	24 / 506 (4.7%)	(4.5%)	(4.3%)	(4.9%)
상장유지를 상장폐지로 예측 (오답률 %)	**14 / 103 (13%)**	(32.0%)	**(19.4%)**	(34.9%)

- **이익변수 추가O 단독모델**

  ▼ 이익변수 추가한 단독모델 n값 조정 별 결과

	term	score	real0_pred1(%)	real1_pred0(%)
0	1년전까지	0.8946	5.749	30.000
1	2년전까지	0.9038	3.673	35.398
2	3년전까지	0.9060	5.317	28.037
3	4년전까지	0.9094	4.990	27.885
4	5년전까지	0.9189	3.205	31.313
5	6년전까지	0.9016	4.232	35.000

	term	score	real0_pred1(%)	real1_pred0(%)
0	1년전까지	0.8738	5.754	45.283
1	2년전까지	0.8705	5.754	47.170
2	3년전까지	0.8754	3.571	54.717
3	4년전까지	0.8721	6.548	42.453
4	5년전까지	0.8852	6.944	33.019
5	6년전까지	0.8787	4.365	49.057
6	7년전까지	0.8607	6.151	50.943

	term	score	real0_pred1(%)	real1_pred0(%)
0	1년전까지	0.8246	0.198	100.000
1	2년전까지	0.8852	7.738	29.245
2	3년전까지	0.8885	7.143	30.189
3	4년전까지	0.8787	5.754	42.453
4	5년전까지	0.8852	3.175	50.943
5	6년전까지	0.8754	6.944	38.679
6	7년전까지	0.8721	6.944	40.566

<단독 모델> * 이익변수 추가, 주식 데이터O	데이터 원형 (n=0)	데이터 원형(n=3)	변화율 (n=4)	지수함수 (n=2)
정확도	0.9295	0.9104	0.9080	0.9064
상장폐지를 상장유지로 예측 (오답률 %)	38 / 504 (7.54%)	(5.389%)	4.941	(5.336%)
상장유지를 상장폐지로 예측 (오답률 %)	**5 / 106 (4.72%)**	**(26.471%)**	(30.097%)	(29.126%)

<단독 모델> * 이익변수 추가, 주식 데이터X	데이터 원형 (n=0)	데이터 원형(n=5)	변화율 (n=5)	지수함수 (n=2)
정확도	0.9311	0.9094	0.8852	0.8852
상장폐지를 상장유지로 예측 (오답률 %)	32 / 504 (6.35%)	(4.99%)	(6.944%)	(7.738%)
상장유지를 상장폐지로 예측 (오답률 %)	**10 / 106 (9.43%)**	(27.885%)	**(33.019%)**	(29.245%)

▼ [결과 비교] 주식 데이터 유무에 따른 응용모델 성능

- 이익 변수 포함이 default
- **{ 데이터 원형(n=0) + 변화율 + 지수함수 } 응용모델이 대체적으로 가장 성능이 좋음**

<응용 모델> * 주식 데이터O	데이터 원형(n=0) + 변화율 (n=4)	데이터 원형(n=0) + 지수함수 (n=3)	데이터 원형(n=0) + 변화율 (n=4) + 지수함수 (n=2)	데이터 원형(n=0) + row로 나열(n=2) + 변화율 (n=3) + 지수함수 (n=2)
정확도	0.9245	0.9278	0.9295	0.9278
상장폐지를 상장유지로 예측	39 / 504 (7.74%)	38 / 504 (7.54%)	39 / 504 (7.74%)	39 / 504 (7.94%)
상장유지를 상장폐지로 예측	**4 / 106 (3.77%)**	6 / 106 (5.66%)	**4 / 106 (3.77%)**	**3 / 106 (2.83%)**

<응용 모델> * 주식 데이터X	데이터 원형(n=0) + 변화율 (n=3)	데이터 원형(n=0) + 지수함수 (n=3)	데이터 원형(n=0) + 변화율 (n=3) + 지수함수 (n=2)	데이터 원형(n=0) + 변화율 (n=5) + 지수함수 (n=2)
정확도	0.9262	0.9311	0.9245	0.9311
상장폐지를 상장유지로 예측	39 / 504 (7.74%)	38 / 504 (7.54%)	42 / 504 (8.33%)	35 / 504 (6.94%)
상장유지를 상장폐지로 예측	6 / 106 (5.66%)	6 / 106 (6.6%)	**4 / 106 (3.77%)**	7 / 106 (6.6%)

▼ (1) 데이터 원형(n=0) + 변화율 (n=3)

> X_t + { $(X_t - X_{t-1})/(X_{t-1})$ ... $(X_t - X_{t-2})/(X_{t-2})$ ... $(X_t - X_{t-n})/(X_{t-n})$ }
>
> - t시점과 $t-n$ 시점 변화율: $(X_t - X_{t-1})/(X_{t-1})$

- **학습**
  - **독립변수** `['cash/assets':'retained_earning']`, `['year']`
  - **종속변수** `['fnc_rsn_unlst_year']`

X_t 시점 재무정보 데이터 원형			X_t 과 X_{t-n}시점까지 최소자승법 지수함수 절편값		기업코드	종속변수
interest/ebitda	inven/sales	...	assets_growth	sales_growth	code_label	fnc_rsn_unlst_year
0.000000	0.148010	...	66.432163	44.816028	5930	0
0.000000	0.218229	...	66.432163	44.816028	373220	0
0.011269	0.207374	...	59.469070	39.641068	660	0
0.021037	0.645961	...	11.197272	34.842612	207940	0
0.024660	0.183522	...	26.693612	12.407985	6400	0
...	...	...	...	...	...	...
0.000000	0.161454	...	0.707426	-0.187998	900180	1
0.410927	0.372227	...	3.411914	-0.286480	950010	1
-0.011492	0.045593	...	-0.181762	-0.466248	950030	1
-0.000000	0.085545	...	0.622113	-0.734431	950070	1
0.018753	0.007705	...	-0.166592	-0.313956	950180	0

- **결과**

y_real	y_pred	
0	0	465
	1	39
1	0	6
	1	100

- 상장폐지를 상장유지로 예측: 39 / 504
- 상장유지를 상장폐지로 예측: 6 / 106

▼ (2) 데이터 원형(n=0) + 지수함수 (n=2)

> X_t + { $a * \exp(bX_t) \dots a * \exp(bX_{t-1}) \dots a * \exp(bX_n)$ }
>
> - t시점과 $t-n$ 시점까지 최소자승법 기반 지수함수 파라미터: $a * \exp(bX)$

- **학습**
  - **독립변수** `['cash/assets':'retained_earning']`, `['year']`
  - **종속변수** `['fnc_rsn_unlst_year']`

X_t 시점 재무정보 데이터 원형			X_t 과 X_{t-n}시점까지 최소자승법 지수함수 파라미터값		기업코드	종속변수
interest/ebitda	inven/sales	...	assets_growth	sales_growth	code_label	fnc_rsn_unlst_year
0.000000	0.148010	...	66.432163	44.816028	5930	0
0.000000	0.218229	...	66.432163	44.816028	373220	0
0.011269	0.207374	...	59.469070	39.641068	660	0
0.021037	0.645961	...	11.197272	34.842612	207940	0
0.024660	0.183522	...	26.693612	12.407985	6400	0
...	...	...	...	...	...	...
0.000000	0.161454	...	0.707426	-0.187998	900180	1
0.410927	0.372227	...	3.411914	-0.286480	950010	1
-0.011492	0.045593	...	-0.181762	-0.466248	950030	1
-0.000000	0.085545	...	0.622113	-0.734431	950070	1
0.018753	0.007705	...	-0.166592	-0.313956	950180	0

- **결과**

y_real	y_pred	
0	0	466
	1	38
1	0	7
	1	99

- 상장폐지를 상장유지로 예측: 38 / 504
- 상장유지를 상장폐지로 예측: 7 / 106

▼ (3) 데이터 원형(n=0) + 변화율 (n=3) + 지수함수 (n=2)

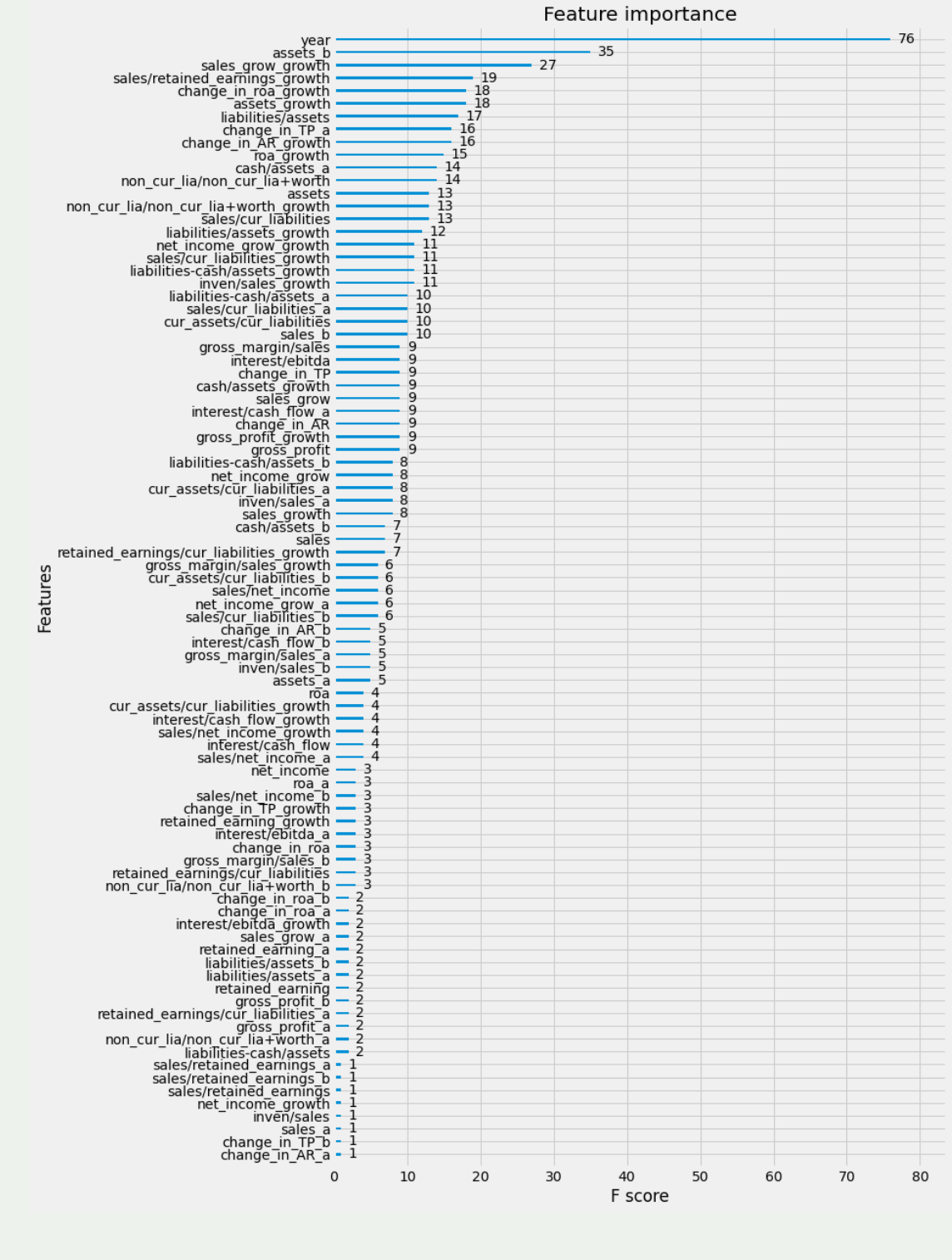
- **학습**
  - **독립변수** `['cash/assets':'retained_earning']`, `['year']`
  - **종속변수** `['fnc_rsn_unlst_year']`

X_t 시점 재무정보 데이터 원형		X_t 과 X_{t-1} 시점 재무정보 변화율		X_t 과 X_{t-n} 시점까지 최소자승법 지수함수 기울기값&절편값		기업코드	종속변수
retained_earning	...	interest/ebitda_growth	...	net_income_b	net_income_grow_a	code_label	fnc_rsn_unlst_year
2.930648e+11	...	-1.000000	...	0.412902	21.48	5930	0
3.375870e+08	...	-1.000000	...	NaN	NaN	373220	0
5.578407e+10	...	84.975640	...	0.703428	136.87	660	0
2.347187e+09	...	-2.726036	...	0.490616	18.76	207940	0
8.516473e+09	...	inf	...	0.683968	56.81	6400	0
...	...	...	...	...	...	...	...
2.412923e+08	...	NaN	...	-0.147541	NaN	900180	1
-5.264892e+07	...	inf	...	NaN	NaN	950010	1
-1.131896e+07	...	-37.637030	...	NaN	NaN	950030	1
-9.833684e+07	...	NaN	...	NaN	NaN	950070	1
3.271038e+07	...	inf	...	NaN	NaN	950180	0

- **결과**

y_real	y_pred	
0	0	462
	1	42
1	0	4
	1	102

- 상장폐지를 상장유지로 예측: 42 / 504
- 상장유지를 상장폐지로 예측: 4 / 106



▼ (4) 데이터 원형(n=0) + row로 나열(n=2) + 변화율 (n=3) + 지수함수 (n=2)

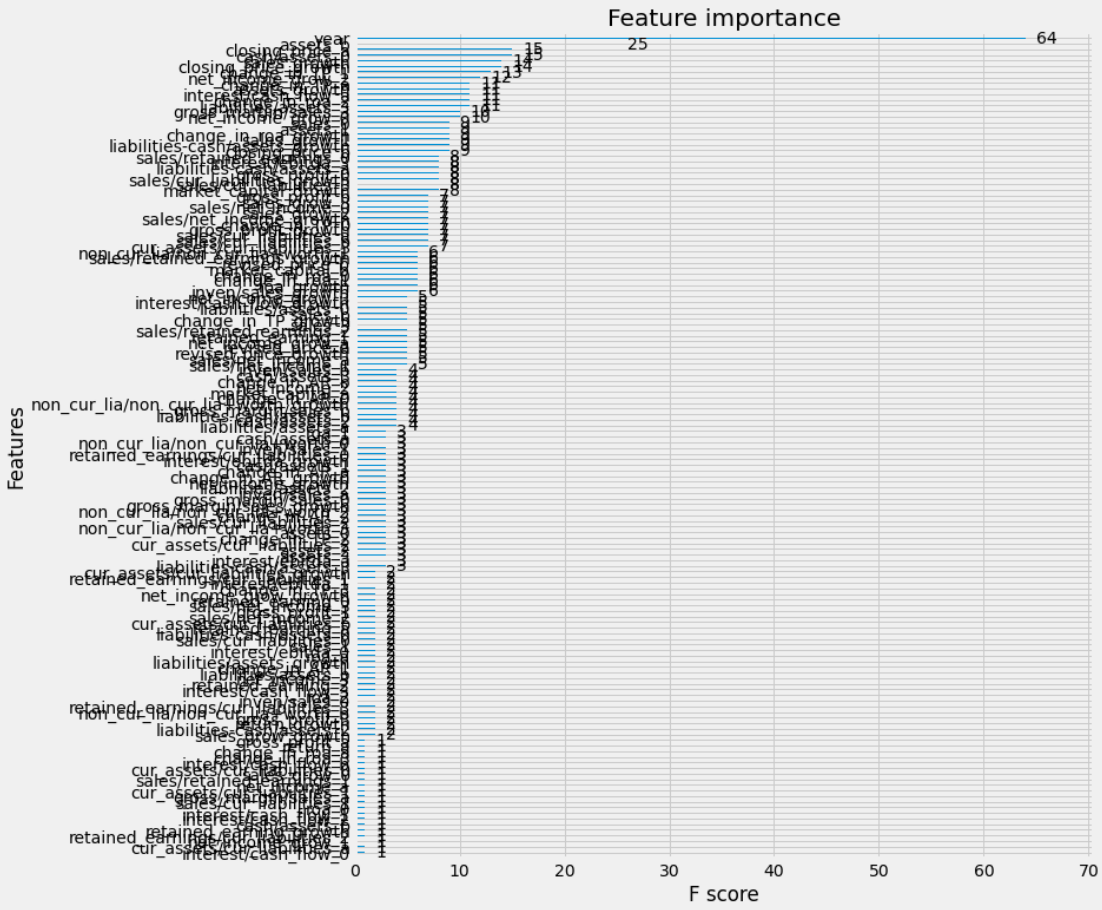
- **학습**
  - **독립변수** `['cash/assets':'retained_earning']`, `['year']`
  - **종속변수** `['fnc_rsn_unlst_year']`

X_t ...X_{t-n} 시점 재무정보 데이터 원형(row로 과거정보 나열)			X_t 과 X_{t-1}시점 재무정보 변화율		X_t 과 X_{t-n}시점까지 최소자승법 지수함수 기울기값&절편값		기업코드	종속변수
interest/ebitda_0	..	retained_earning	...	interest/ebitda_growth	...	net_income_b	net_income_grow_a	code_label	fnc_rsn_unlst_year
0.000000	...	2.930648e+11	...	-1.000000	...	0.412902	21.48	5930	0
0.052709	...	3.375870e+08	...	-1.000000	...	NaN	NaN	373220	0
0.107330	...	5.578407e+10	...	84.975640	...	0.703428	136.87	660	0
0.181923	...	2.347187e+09	...	-2.726036	...	0.490616	18.76	207940	0
0.010386	...	8.516473e+09	...	inf	...	0.683968	56.81	6400	0
...	...	...	...	...	...	...	...	...	...
-0.825398	...	2.412923e+08	...	NaN	...	-0.147541	NaN	900180	1
0.070275	...	-5.264892e+07	...	inf	...	NaN	NaN	950010	1
-0.080851	...	-1.131896e+07	...	-37.637030	...	NaN	NaN	950030	1
0.012212	...	-9.833684e+07	...	NaN	...	NaN	NaN	950070	1
0.167297	...	3.271038e+07	...	inf	...	NaN	NaN	950180	0

- **결과**

y_real	y_pred	
0	0	465
	1	39
1	0	3
	1	103

- 상장폐지를 상장유지로 예측: 39 / 504
- 상장유지를 상장폐지로 예측: 3 / 106



▼ 4. 응용 모델 & 모델 별로 주식 데이터 유무

<응용 모델> * 이익변수 추가 *모델 별로 주식 데이터 유무 다르게 적용	데이터 원형(n=0, 주식O) + 지수함수 (n=3, 주식X)			

<응용 모델> * 이익변수 추가 *모델 별로 주식 데이터 유무 다르게 적용	데이터 원형(n=0, 주식O) + 지수함수 (n=3, 주식X)			
정확도	0.93114			
상장폐지를 상장유지로 예측	38 / 504 (7.54%)			
상장유지를 상장폐지로 예측	4 / 106 (3.77%)			

📌 **결론**

- **{ 데이터 원형(n=0) + 변화율 + 지수함수 }** 응용모델이 대체적으로 가장 성능이 좋음
- 현재까지 실험에서 '**상장유지(0)를 상장폐지(1)로 예측'** 오답 개수를 최소로 줄이면, **3 / 106 (2.83%)**
- '**상장유지(0)를 상장폐지(1)로 예측'** 오답 개수는 3~4개로 고정하되, '**상장폐지(1)를 상장유지(0)로 예측'** 오답개수를 더 줄이는 최적의 조합을 찾는 것을 목표로 할 것임.
- `'year'` 변수에 대한 의존도가 큼. 특히 응용모델인 경우 단독모델일 때보다 의존도 더 큼. → 과적합 의심

- 단독모델(원데이터, 변화율, 지수함수)을 조합한 **응용모델을 만들 때**,
  - 각 단독모델 별 주식 데이터를 포함유무을 달리하고 조합하면 성능이 개선됨
    ex. 변화율 모델은 주식데이터를 포함하지 않고, 지수함수 모델은 주식데이터를 포함 시킨 후 조합해 응용모델은 만들면 성능 개선
  - 단독 모델 최적 n과 응용모델로 조합후 최적 n이 다름
    즉, 조합 후 최적 n값을 다시 찾아야 함