Final Project Team 1

Bankruptcy

김민선 박중창 오다건 오재욱 이청파

Last week

01 Missing Value delete & impute

02 Data type change

03 Scaling

04 Attribute research

05 Simple modeling

Data

01 재무회계표, 현금흐름표, 손익계산서로 이루어진 재무제표

1. 여러 재무비율들의 항목들
2. 다만 정확히 무슨 항목이 어떤 카테고리에 들어가는지는 파악할 수 없었음
3. High - correlated
4. Selected Variables 파트에서 다시 !

!! Dist'n, Multicollinearity

01 Dist'n : Skewed ➡️ RobustScaler

02 Multicollinearity

1. [방법1] VIF & Corr 0.9 이상
2. [방법2] Item 별 Clustering 후 Corr 0.9 이상 변수 제거를 합집한 후 Corr 0.9 이상의 변수들 삭제.
3. Pre-processed 된 데이터에 총 42개의 변수

'Attr1', 'Attr2', 'Attr3', 'Attr4', 'Attr5', 'Attr6', 'Attr7', 'Attr8', 'Attr9', 'Attr10', 'Attr12', 'Attr13', 'Attr15', 'Attr18', 'Attr19', 'Attr20', 'Attr21', 'Attr24', 'Attr25', 'Attr27', 'Attr28', 'Attr29', 'Attr30', 'Attr32', 'Attr33', 'Attr38', 'Attr39', 'Attr40', 'Attr41', 'Attr42', 'Attr43', 'Attr45', 'Attr47', 'Attr49', 'Attr51', 'Attr53', 'Attr55', 'Attr57', 'Attr58', 'Attr59', 'Attr61', 'Attr64'

!! Selected Variables

01 Common items in Left Variables

재무상태표	손익계산서	현금흐름표
Total Assets Short-term liabilities Total Liabilities Equity Receivables Inventory Working capital Fixed Assets….	Gross Profit Sales	Depriciation

Modeling plan

01 Base Model

- Full data
- With preprocessed Data

02 grouping+PCA

이때 grouping 이란, 변수들을 특정 집단으로 묶는것

03 grouping + PCA + consider correlation

Initial model

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9626	0.8914	0.3478	0.8000	0.4848	0.4685	0.5128
1	0.9714	0.9300	0.4348	1.0000	0.6061	0.5936	0.6497
2	0.9692	0.8755	0.4348	0.9091	0.5882	0.5743	0.6168
3	0.9670	0.9270	0.5217	0.7500	0.6154	0.5987	0.6095
4	0.9736	0.9391	0.5417	0.9286	0.6842	0.6714	0.6981
5	0.9735	0.9223	0.6087	0.8235	0.7000	0.6865	0.6951
6	0.9757	0.9446	0.6087	0.8750	0.7179	0.7057	0.7184
7	0.9647	0.9088	0.3913	0.8182	0.5294	0.5134	0.5515
8	0.9713	0.9496	0.4783	0.9167	0.6286	0.6152	0.6506
9	0.9625	0.8595	0.3478	0.8000	0.4848	0.4685	0.5128
Mean	0.9691	0.9148	0.4716	0.8621	0.6040	0.5896	0.6215
SD	0.0045	0.0288	0.0921	0.0724	0.0797	0.0810	0.0714

Prediction of Unseen Data
0.5

Base model

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9559	0.8717	0.3043	0.6364	0.4118	0.3918	0.4208
1	0.9714	0.9237	0.4348	1.0000	0.6061	0.5936	0.6497
2	0.9581	0.9192	0.3043	0.7000	0.4242	0.4060	0.4444
3	0.9758	0.8976	0.5217	1.0000	0.6857	0.6744	0.7133
4	0.9670	0.9140	0.4583	0.8462	0.5946	0.5790	0.6087
5	0.9735	0.9430	0.6522	0.7895	0.7143	0.7005	0.7040
6	0.9625	0.9270	0.5217	0.6667	0.5854	0.5660	0.5707
7	0.9669	0.9070	0.4348	0.8333	0.5714	0.5560	0.5880
8	0.9625	0.9182	0.3478	0.8000	0.4848	0.4685	0.5128
9	0.9558	0.8417	0.3043	0.6364	0.4118	0.3918	0.4208
Mean	0.9649	0.9063	0.4284	0.7908	0.5490	0.5328	0.5633
SD	0.0068	0.0281	0.1101	0.1280	0.1051	0.1075	0.1048

Prediction of Unseen Data
0.4375

Grouping + PCA 2

01 Grouping

Group1 : short term liabilities 관련 변수들
Group2 : long term liabilities 관련 변수들
Group3 : others

02 PCA

각 Group 에서 3,1,3개의 PC 추출

03 Mess !

Table 3: Ratios in Model#2

Ratios
book value of equity / total liabilities
equity / total assets
gross profit / short-term liabilities
(inventory * 365) / sales
operating expenses / short-term liabilities
(current assets - inventory - receivables) / short-term
profit on operating activities / sales
(current assets - inventory) / short-term liabilit
EBITDA (profit on operating activities - depreciatior
long-term liabilities / equity
sales / short-term liabilities
sales / fixed assets

Table 2: Ratios in Model#1

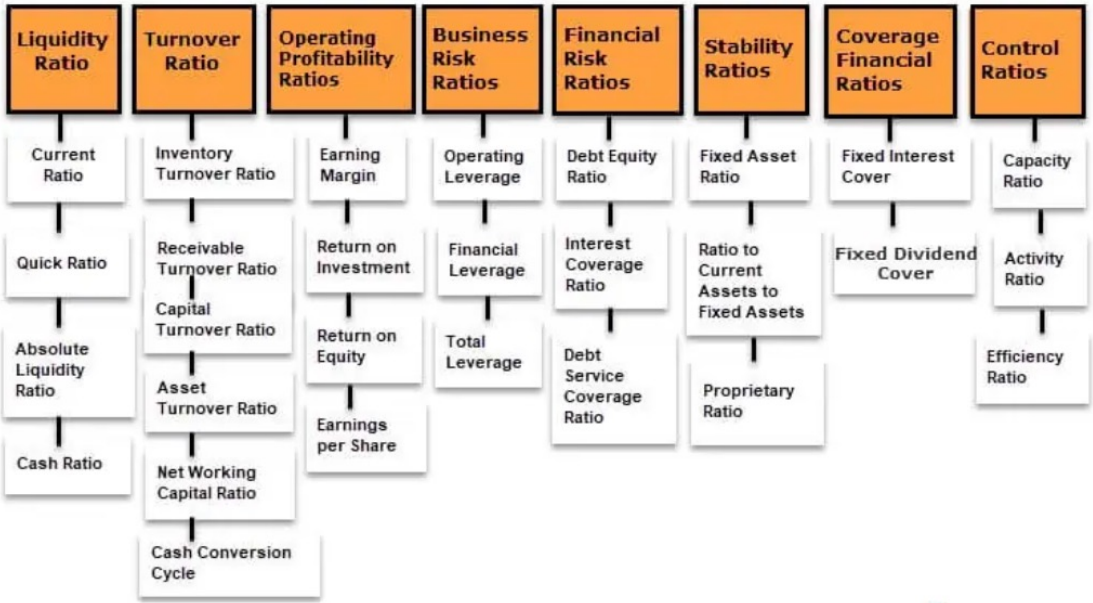
Ratios
[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365
gross profit (in 3 years) / total assets
(equity - share capital) / total assets
(net profit + depreciation) / total liabilities
operating expenses / total liabilities

Mukeri, et al. (2020). Financial Data Analysis Using Expert Bayesian Framework For Bankruptcy Prediction. arXiv e-prints. 7-8

Grouping + PCA 2

01 Grouping

특정 재무비율군의 의미만 맞으면 해당 카테고리로 분류 가능하여 조사해서 분류



WallStreetMojo

PREDICTING CORPORATE TURNAROUND USING LOGISTIC REGRESSION ANALYSIS: A RESEARCH ON BURSA MALAYSIA COMPANIES 218p

Grouping + PCA 2

01 Grouping

Liquidity	Turnover Ratio	Operating profitability	Business Risk	Financial Risk	Stability
12,32,33,4 ,40,5,51	1,15,20,43 ,47,57,61, 64,9	18,21,24, 27,3,41,42 ,45,49,53, 7	13,39,19	10,2,25,30 ,58,59,6,8	28,29,38, 55

Grouping + PCA 2

02 PCA

1. 각 군집 별로 Standard Scaler 이용하여 분산 크기 조정
2. 군집 내에서 **선형성**을 만족하는 변수들을 추려 (corr over 0.5) 해당 변수들에 대해서만 PCA
3. 해당 변수들에 PCA 적용 후 분산의 설명력이 0.7 이상일 경우 cut ,5개 변수 제거

Grouping + PCA 2

03 Modeling

Before Tuning

	Model	Accuracy	AUC	Recall	Prec.	F1
catboost	CatBoost Classifier	0.9679	0.9348	0.3916	0.9212	0.5451
xgboost	Extreme Gradient Boosting	0.9669	0.9247	0.3846	0.8945	0.5346
lightgbm	Light Gradient Boosting Machine	0.9669	0.9211	0.3777	0.8997	0.5295

Tuning…catboost

Mean	0.9693	0.9244	0.4256	0.9182	0.5754	0.5620	0.6098
SD	0.0035	0.0251	0.0841	0.0515	0.0762	0.0765	0.0588

Grouping + PCA 2

03 Modeling

Before Tuning

	Model	Accuracy	AUC	Recall	Prec.	F1
catboost	CatBoost Classifier	0.9679	0.9348	0.3916	0.9212	0.5451
xgboost	Extreme Gradient Boosting	0.9669	0.9247	0.3846	0.8945	0.5346
lightgbm	Light Gradient Boosting Machine	0.9669	0.9211	0.3777	0.8997	0.5295

After Tuning

Mean	0.9693	0.9244	0.4256	0.9182	0.5754	0.5620	0.6098
SD	0.0035	0.0251	0.0841	0.0515	0.0762	0.0765	0.0588

	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC
0	0.9743	0.9167	0.5667	0.8947	0.6939	0.6812	0.7006
1	0.9657	0.8895	0.3103	1.0000	0.4737	0.4610	0.5473
2	0.9691	0.9089	0.4483	0.8667	0.5909	0.5765	0.6106
3	0.9743	0.9209	0.5172	0.9375	0.6667	0.6544	0.6859
4	0.9691	0.9795	0.4828	0.8235	0.6087	0.5938	0.6168

Grouping + PCA 2

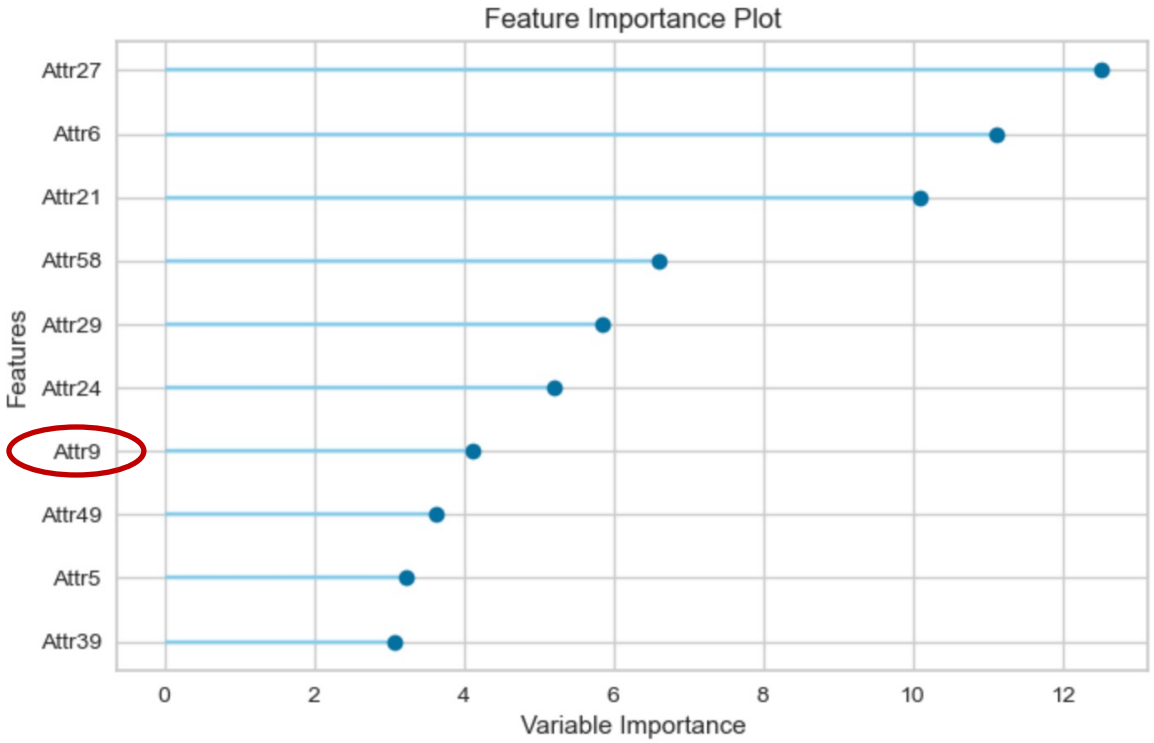
04 Prediction of Unseen data

Light GBM	Catboost
0.63xx	0.6957

Why?

1. Tree 를 분할하는 방법이 다른 두 모델
2. Catboost 는 하이퍼 파라미터 튜닝에 예민한 모델

Feature importance of base model catboost



34가 빠지고 9,48,38 등장!

So, is it predictable?

1. 가능이야 하겠지만…
2. Imbalance 된 데이터로 인해 변수를 줄이면서 최대한의 예측력을 갖추기 쉽지 않음.
3. 또한, 같은 이유로 f1 score 가 적절한 metric인지도 의문스러움.
   - ➡️ ROC curve
   - ➡️ 파산할 기업을 현행유지라고 잘못 판단하는 것이 더 치명적…!

-The END-