Final Project Team 1

Bankruptcy

김민선 박중창 오다건 오재욱 이청파

Pre-Processing

01 Missing Value

1. Convert “?” (NA) to NA

2. 1.4% of data (total 6855 obs)

3. # of NA’s (Variables)

   Attr 37: **45%**
   Attr 25, 45, 60: 6% 👉 **Delete** Attr37

4. Replace NA with **mean**

Pre-Processing

02 Data type

Mixed **Str**, **Numeric (Object)** ➡️ Applied **numeric** to all

03 Duplicates

20 duplicated rows ➡️ drop

04 Convert Y

To categorical variable

EDA

01 Skewness &Outliers

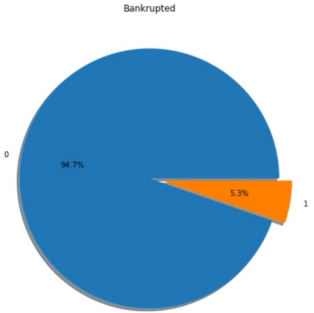
Mostly **highly** Skewed & Lots of outlier ➡️ Apply “**RobustScaler**”

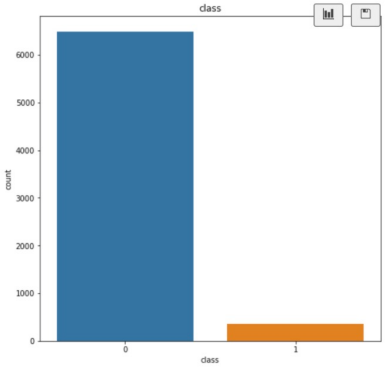
02 Unit

Attr(29, 43, 55) : numeric values / Others : Ratio

03 Imbalanced y value

Value	Count	Frequency (%)
0.0	6494	94.7%
1.0	361	5.3%





EDA

04 64 Attribute research

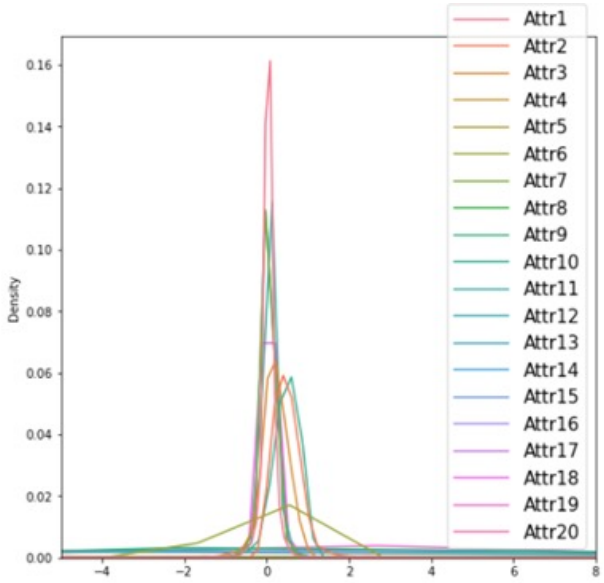
example

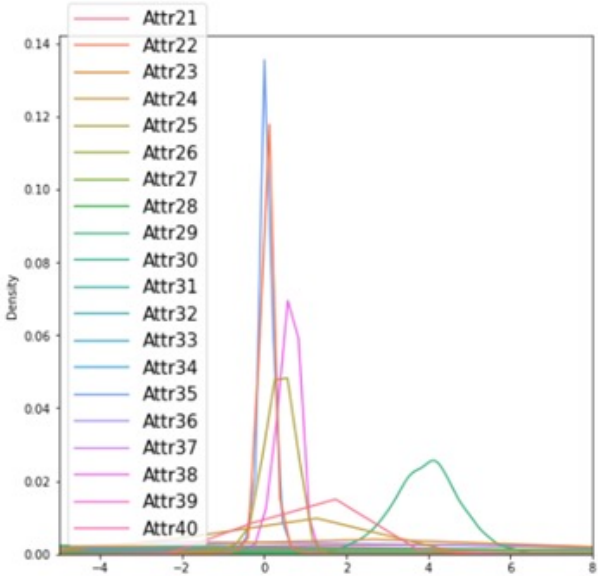
9. `X22` : ROOA(Return on operating assets)
    a. `profit on operating activities / total assets`
        i. profit ~ : 유가증권(주식, 채권 등)과 ***자산***(부동산 등) 을 투자자의 이익을 위하여 정해진 투자목적에 맞게 전문적으로 운용하는 것
    b. **percentage profit** that a company can expect from the purchase of a new piece of equipment

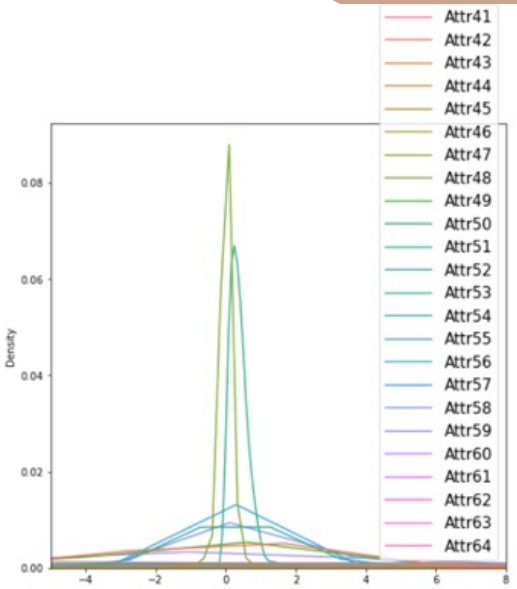
       회사의 자산으로 이익을 얼마나 만들어냈는지

> mean -0.359244
> std 23.935438
> min -1578.700000
> 25% 0.001026
> 50% 0.026098
> 75% 0.072874
> max 497.020000

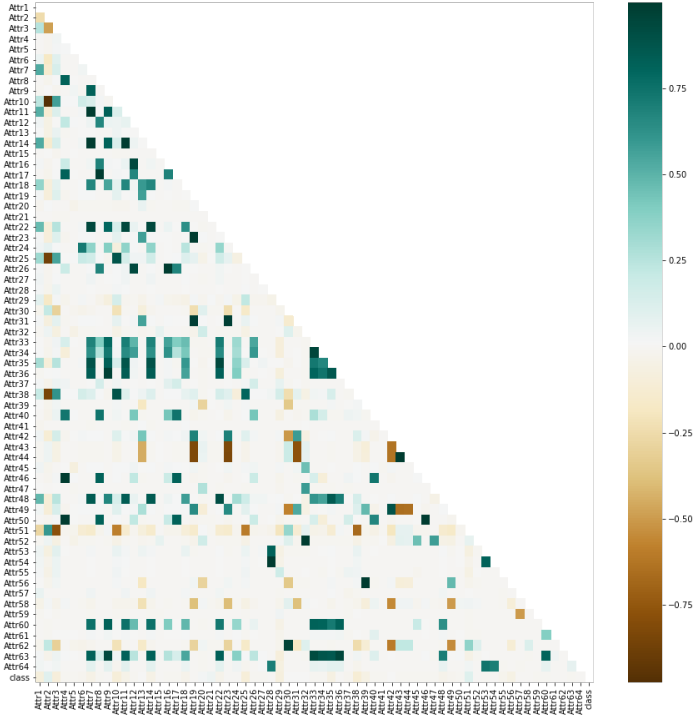
Dist'n of Attr







Correlation heatmap



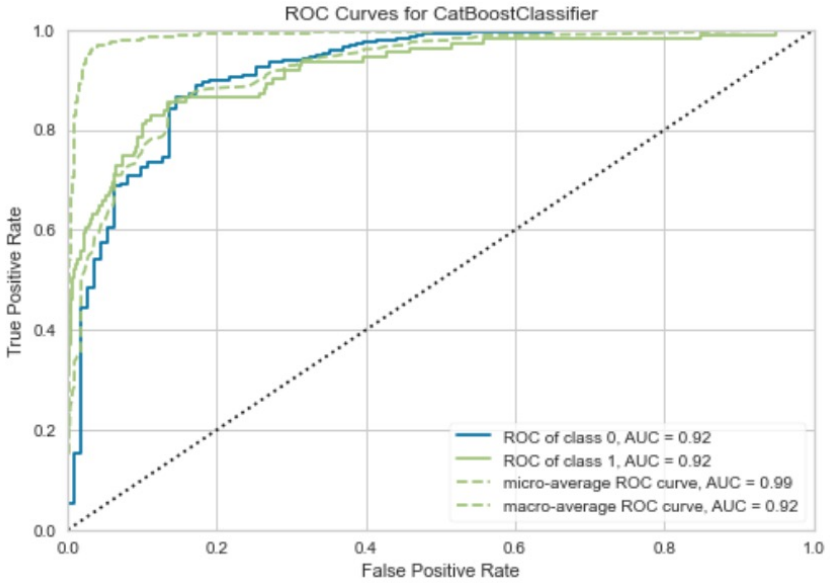
Simple Modelling

MODEL	F1 score
Logistic Regression	0.0235
Xgboost Classifier	0.5494
Light GBM Classifier	0.5194
Random Forest Classifier	0.2067
Catboost Classifier	**0.5803**

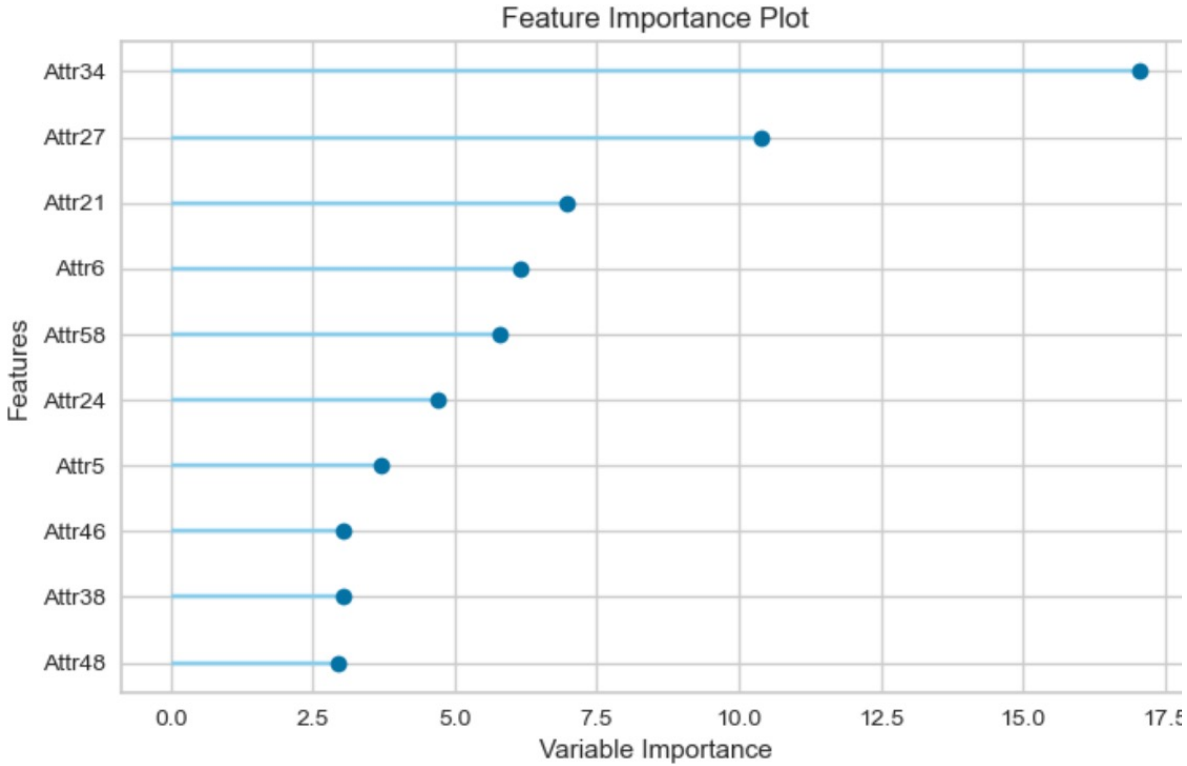
Comparing Models with "Pycaret"

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
catboost	CatBoost Classifier	0.9693	0.9358	0.4237	0.9481	0.5803	0.5670	0.6195	6.6510
xgboost	Extreme Gradient Boosting	0.9665	0.9278	0.4069	0.8766	0.5494	0.5346	0.5805	1.1590
lightgbm	Light Gradient Boosting Machine	0.9656	0.9230	0.3723	0.9087	0.5194	0.5051	0.5639	0.3970
gbc	Gradient Boosting Classifier	0.9638	0.9140	0.3725	0.8329	0.5067	0.4911	0.5380	1.5260
rf	Random Forest Classifier	0.9539	0.8776	0.1216	0.8400	0.2067	0.1969	0.2989	0.4320
ada	Ada Boost Classifier	0.9510	0.8518	0.2034	0.5836	0.2911	0.2722	0.3176	0.3350
et	Extra Trees Classifier	0.9480	0.8412	0.0174	0.1500	0.0308	0.0261	0.0433	0.1670
ridge	Ridge Classifier	0.9471	0.0000	0.0043	0.1000	0.0083	0.0039	0.0122	0.0100
lr	Logistic Regression	0.9466	0.5792	0.0130	0.1700	0.0235	0.0170	0.0341	1.1800
knn	K Neighbors Classifier	0.9466	0.6185	0.0174	0.2333	0.0320	0.0246	0.0502	0.0390
lda	Linear Discriminant Analysis	0.9422	0.7253	0.0435	0.2127	0.0699	0.0530	0.0725	0.0170
dt	Decision Tree Classifier	0.9347	0.6851	0.4071	0.3843	0.3934	0.3592	0.3603	0.0680
svm	SVM - Linear Kernel	0.8465	0.0000	0.2388	0.0905	0.1255	0.0647	0.0744	0.0120
qda	Quadratic Discriminant Analysis	0.5451	0.5270	0.5150	0.0575	0.1034	0.0130	0.0273	0.0160
nb	Naive Bayes	0.1030	0.4968	0.9091	0.0493	0.0936	-0.0034	-0.0283	0.0090

Performance of catboost



Feature importance using catboost



Feature importance using catboost

Attr 34		operating expenses / total liabilities	영업 비용/ 총 부채
Attr 27		profit on operating activities / financial expenses	영업 이익/ 재무 비용
Attr 21	매출액증가율	sales (n) /sales (n-1)	전년도 대비 매출 성장률
Attr 6		retained earnings / total assets	**성장성 지표**. 기업이 이익을 배당하는지 ,혹은 재투자 하는 지에 대한 지표.
Attr 24	3년 총자산 회전율	Gross profit (in 3 years) / total assets	**총자산**에 비해서 매출액을 몇 배나 창출하는지의 지표로, **기업의 활동성**을 나타내는 지표.
Attr 5		[(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - depreciation)] * 365	
Attr 46	당좌 비율	(current assets - inventory) /short-term liabilities	**현금화 할 수 있는 자산의 비율**. 1 넘을수록 안정적이지만 너무 높으면 미래 수익성을 떨어뜨리는 요인
Attr 58		total costs /total sales	총 비용/ 총 매출

Upcoming week

1. Discuss about Feature extraction & handle multicollinearity.
2. Final Modeling with instructions on ESC notion.

-The END-