ESC Final 중간 발표

Choose Dataset

기업 파산 확률

```
X1 net profit / total assets
X2 total liabilities / total assets
X3 working capital / total assets
X4 current assets / short-term liabilities
X5 [(cash + short-term securities + receivables - short-term liabilities) / (operating expenses - deprec
iation)] * 365
X6 retained earnings / total assets
X7 EBIT / total assets
X8 book value of equity / total liabilities

...

class bankruptcy
```

2년 후 기업의 파산 확률 예측하기

→ Classification Model!

Pre-Processing

Missing

- `Attr37` 열 제거 (3100 NA's)
- 열평균으로 NA Imputation

Outliers

- **[방안 1]** 1.5 IQR 규칙을 이용한 이상치 제거
- **[방안 2]** 히스토그램을 그려보며 이상치 판단, 제거

  → 지나친 데이터 손실 방지하기 위해 **[방안 2]** 선택!

Scaling

- `min-max scaler`를 이용한 변수 크기 조정

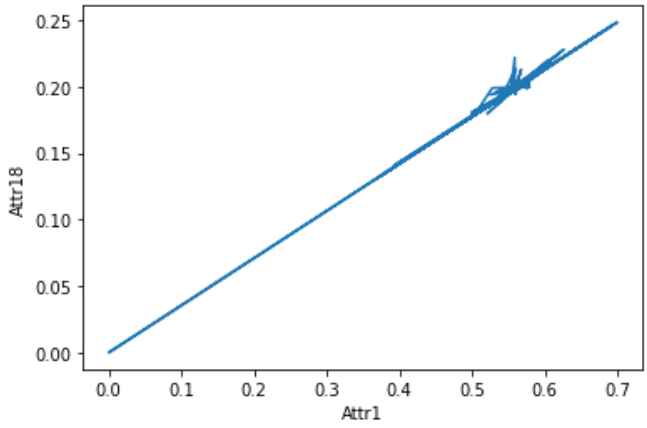
Feature Extraction

Data Reduction

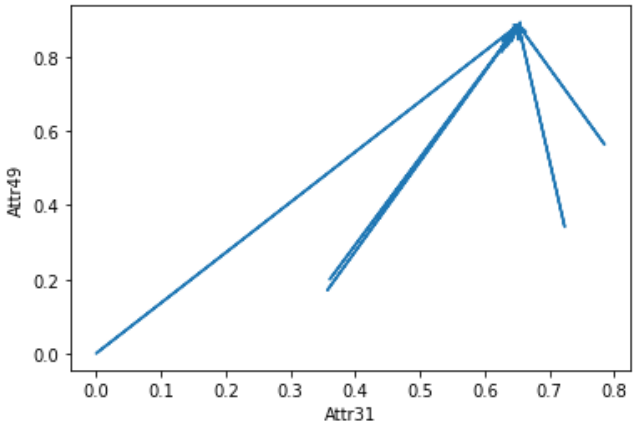
- 일부 변수는 다른 변수와 **역수 관계**에 해당함
  - `Attr2` = total liabilities / total assets
    `Attr17` = total assets / total liabilities

- `Attr20` = (inventory * 365) / sales
  `Attr60` = sales / inventory

- 일부 변수는 다른 변수와 **선형 관계**가 있음!
  - `Attr1` = net profit / total assets
    `Attr18` = gross profit / total assets



  - `Attr13` = (gross profit + depreciation) / sales
    `Attr31` = (gross profit + interest) / sales



  → 상관계수 행렬, VIF 값을 살펴보며 독립적인 설명변수 추출

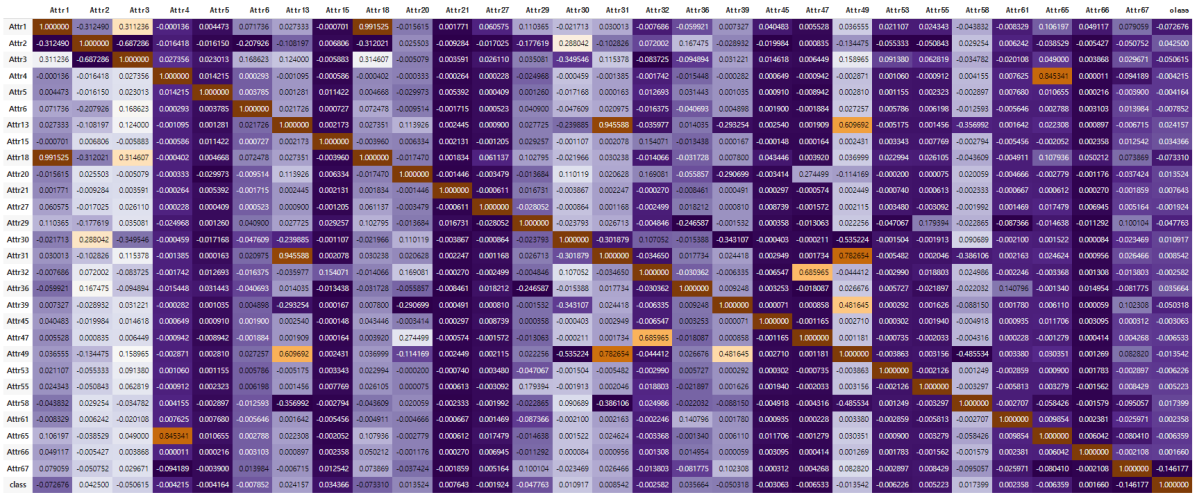
Final Dataset

- 총 26개 변수
  `Attr1`, `Attr2`, `Attr3`, `Attr4`, `Attr5`, `Attr6`, `Attr13`, `Attr15`, `Attr20`, `Attr21`, `Attr27`, `Attr29`, `Attr30`, `Attr32`, `Attr36`, `Attr39`, `Attr45`, `Attr47`, `Attr49` , `Attr53`, `Attr55`, `Attr58`, `Attr61`, `Attr65`, `Attr66`, `Attr67`
- `data.head()`

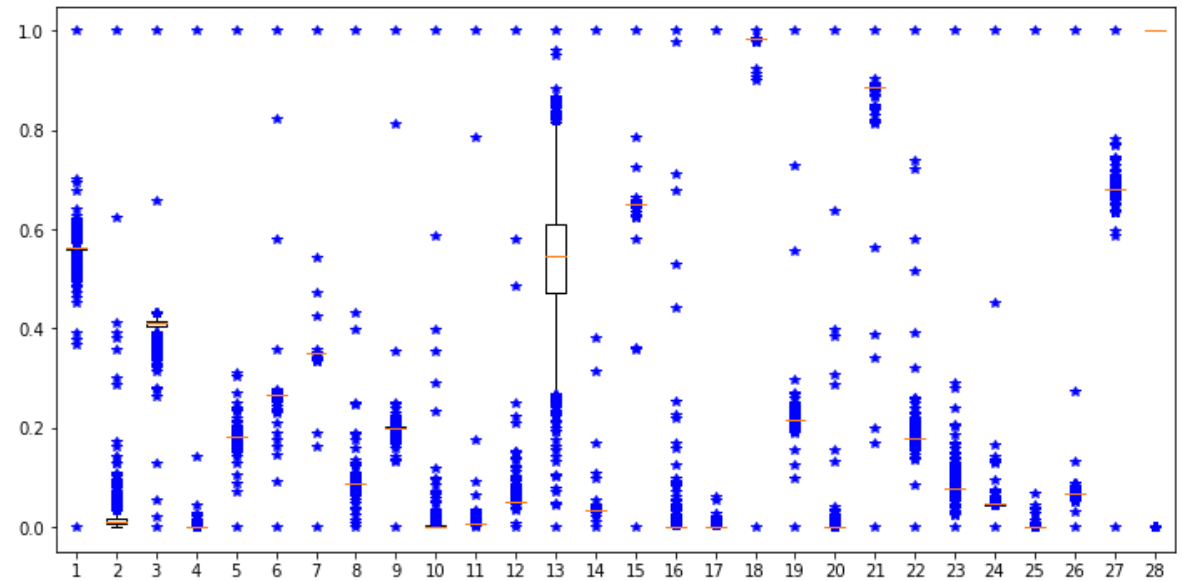
Attr27	Attr29	Attr30	Attr32	Attr36	Attr39	Attr45	Attr47	Attr49	Attr53	Attr55	Attr58	Attr61	Attr65	Attr66	Attr67	class
0.05093548	0.6184205	0.03454141	0.000523037	0.000407364	0.9853279	0.2151856	0.000228313	0.8860730	0.1778310	0.07671939	0.04549398	0.0000342000	0.06591070	0.6796953	1	0
0.05154660	0.1983487	0.03452696	0.000000000	0.000656097	0.9854153	0.2155379	0.000025800	0.8868370	0.1787013	0.07612784	0.04461159	0.0003508250	0.06624784	0.6806493	0	0
0.05091811	0.4183941	0.03453722	0.000541799	0.000890787	0.9852881	0.2155379	0.000025800	0.8857663	0.1779059	0.07615262	0.04599367	0.0000238000	0.06560043	0.6785558	1	0
0.05093589	0.7301522	0.03452106	0.000321179	0.000951440	0.9853381	0.2152740	0.000352241	0.8861732	0.1784206	0.10336769	0.04542307	0.0000953000	0.06618998	0.6799549	1	0
0.05101603	0.9618471	0.03453798	0.000895767	0.000471682	0.9853864	0.2153958	0.000528965	0.8864511	0.1778838	0.07612572	0.04504475	0.0000617000	0.06625422	0.6801929	1	0
0.05093541	0.5560512	0.03453494	0.000767437	0.001123601	0.9853268	0.2152864	0.000330304	0.8861057	0.1778551	0.07615862	0.04532552	0.0000610000	0.06606042	0.6801532	1	0
0.05093679	0.6799352	0.03452782	0.000353869	0.000893720	0.9853313	0.2152361	0.000405974	0.8860849	0.1780178	0.08331362	0.04554972	0.0000759000	0.06599355	0.6798154	1	0

EDA and Visualization

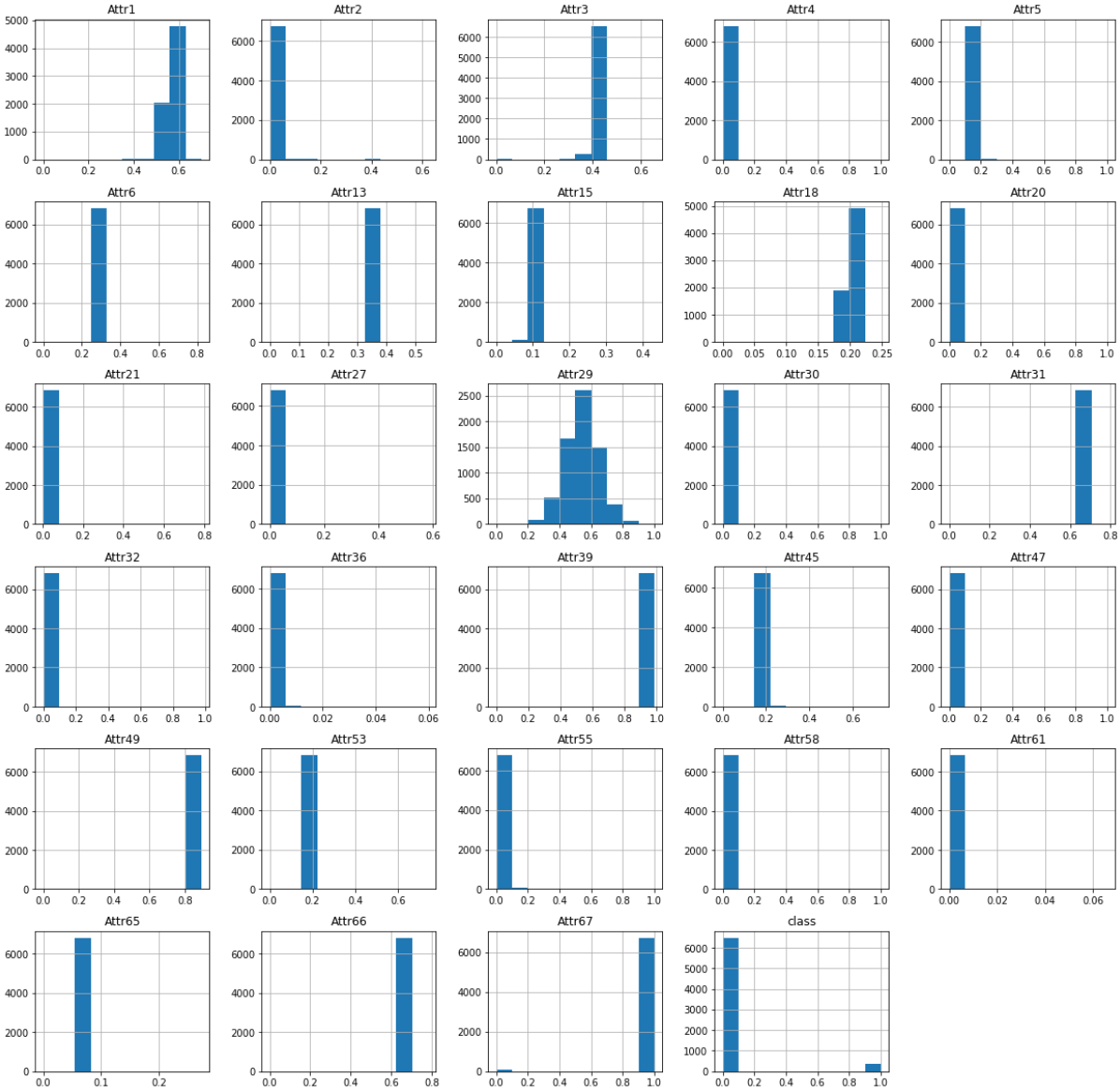
Correlation Plot



이상치 제거



Histogram



Insights

- **Domain Knowledge**

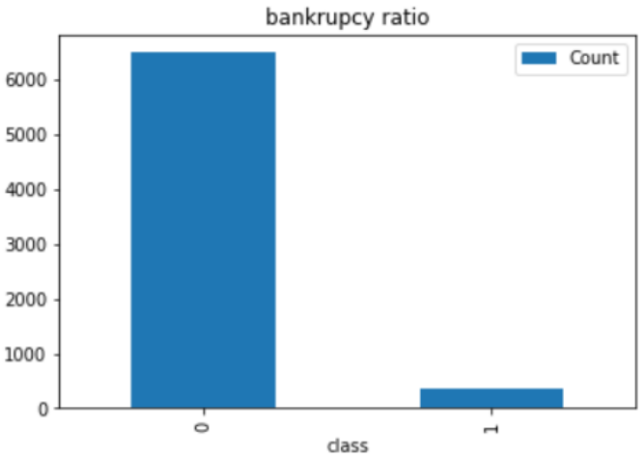
  회계자료를 정확하게 이해하기 위해 기초적인 회계지식이 필수적

  데이터를 보기 전에 변수간 관계를 이해하는 데에 도움이 됨!

- **Skewness**

  일부 설명변수 뿐만 아니라, 종속 변수 역시 상당히 왜도가 심하다!

  비파산 데이터 **(0)** >>> 파산 데이터 **(1)**



→ Imbalanced data에 적합한 접근 방식 고려해 볼 것! (Under/Oversampling 등)

- **Model Candidates**

  ✔️ Logistic Regression

  □ LDA / QDA

  ✔️ SVM

  □ CART / C4.5

  ✔️ Boosting (Adaboost / GBM)

  ✔️ Bagging / Random Forest

  → 마지막으로 **Stacking** 기법을 활용해 합의점을 찾을 예정!

Limitations

- 부도 확률 예측과 더불어 여신 여부에 대한 판단도 고민할 것!