Company Bankruptcy Prediction

Predicting bankruptcy risk through

Machine Learning Classification

Cindy Su - 10/29/2021

Introduction

- **Goal: Predicting bankruptcy risk and possible financial distress of public companies through financial ratios.**

- **User: Bank Creditors or Investors**

- **Product: Machine Learning Classification model for bankruptcy risk prediction**

Data & Methodology

STEP 01 Data collection
STEP 02 Data preparation
STEP 03 Choosing a model
STEP 04 Training
STEP 05 Evaluation
STEP 06 Parameter tuning
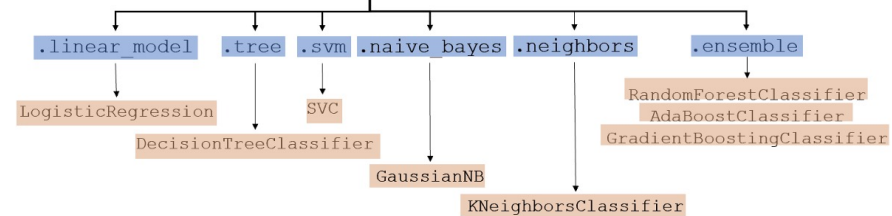**STEP 07** Prediction

kaggle

TAIWAN STOCK EXCHANGE

- Taiwan Economic Journal
- 800-900 listed companies
- From 1999 to 2009
- rows: 6819
- features: 96

pandas

NumPy

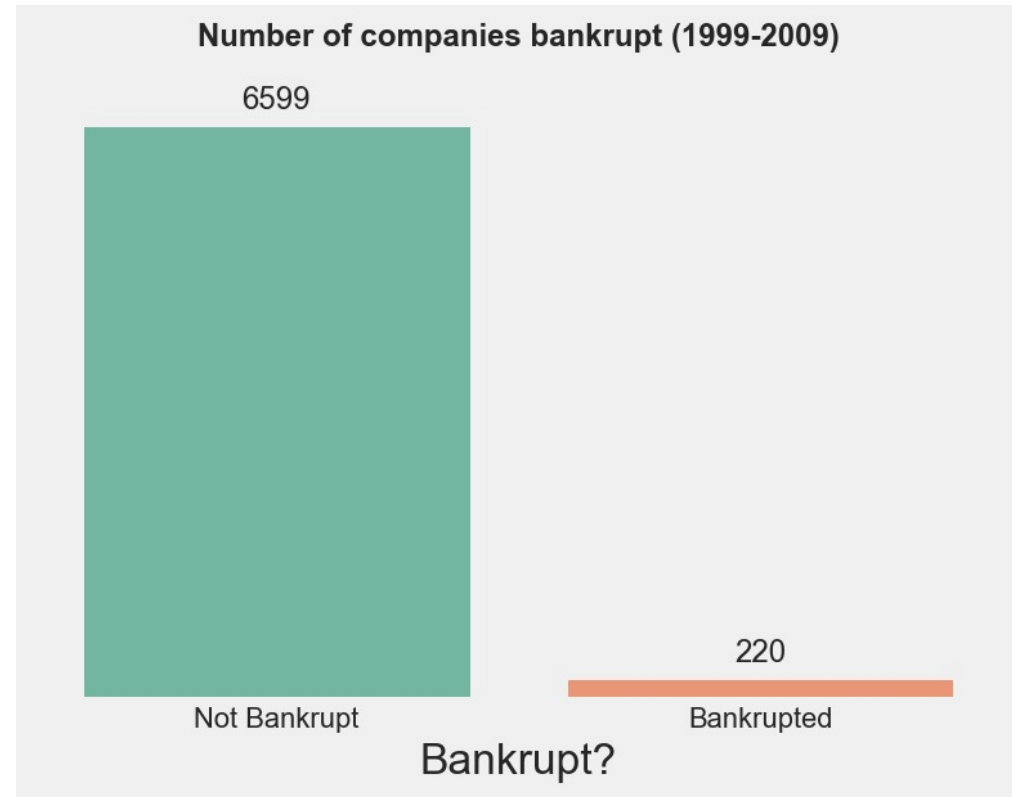- Feature selection & Dimensionality reduction

scikit learn classifiers

- .linear_model
  - LogisticRegression
- .tree
  - DecisionTreeClassifier
- .svm
  - SVC
- .naive_bayes
  - GaussianNB
- .neighbors
  - KNeighborsClassifier
- .ensemble
  - RandomForestClassifier
  - AdaBoostClassifier
  - GradientBoostingClassifier

MERGENT Online

- S&P500
- From 2018-2020

Exploratory Data Analysis

Highly Imbalanced Data

- SMOTE Oversampling

Number of companies bankrupt (1999-2009)
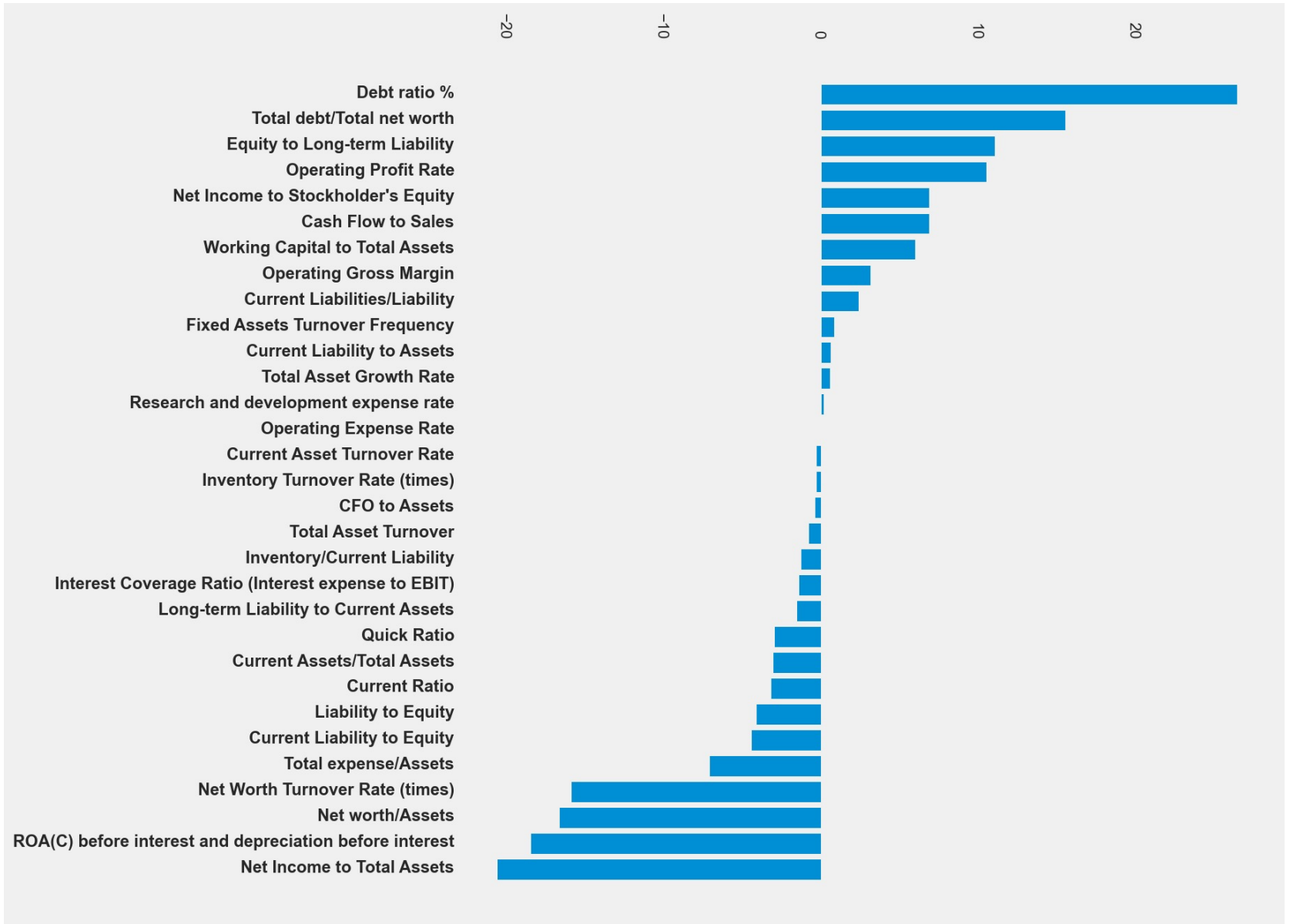
6599

220

Not Bankrupt · Bankrupted

Bankrupt?

- True Class: Bankrupted
- False Class: No bankruptcy
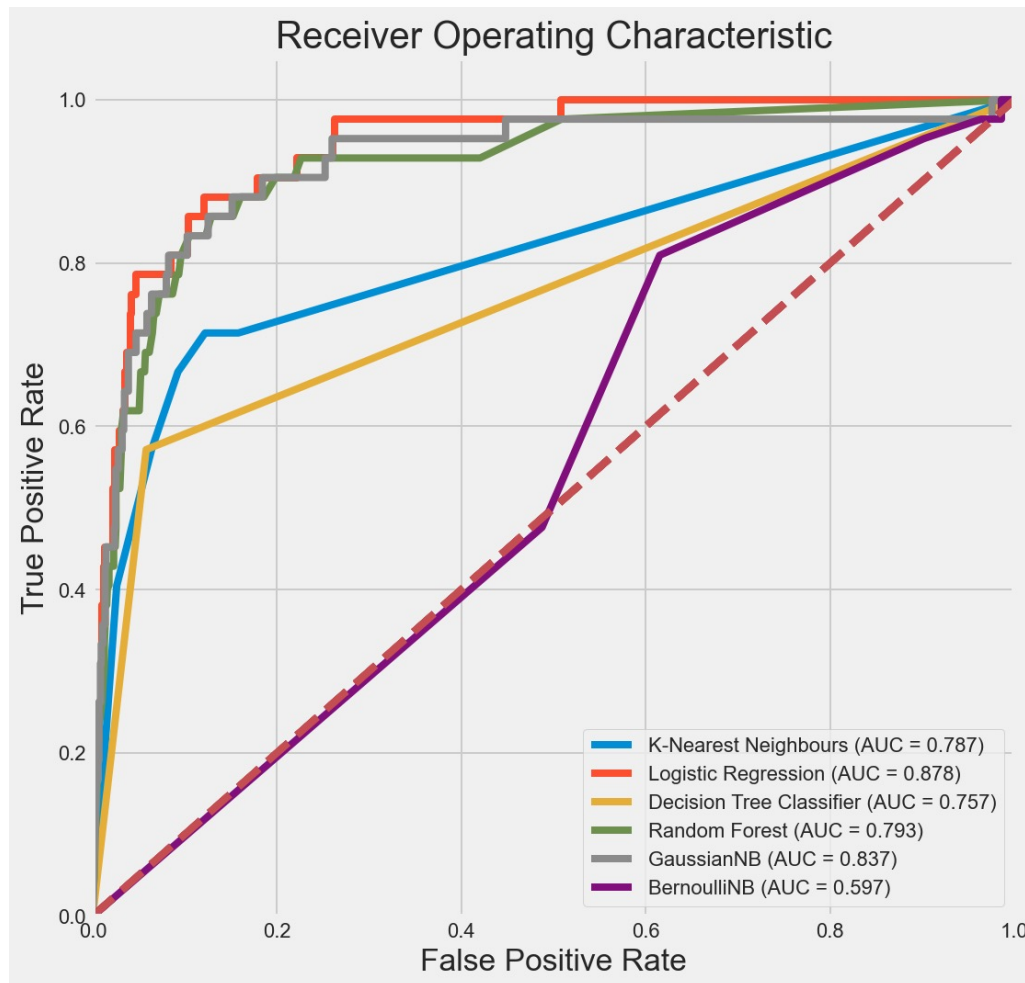- Goal: Minimize False Negative ➡ Maximize Recall

$$Recall = \frac{TP}{TP + FN}$$

Exploratory Data Analysis

Feature selection & Dimensionality reduction

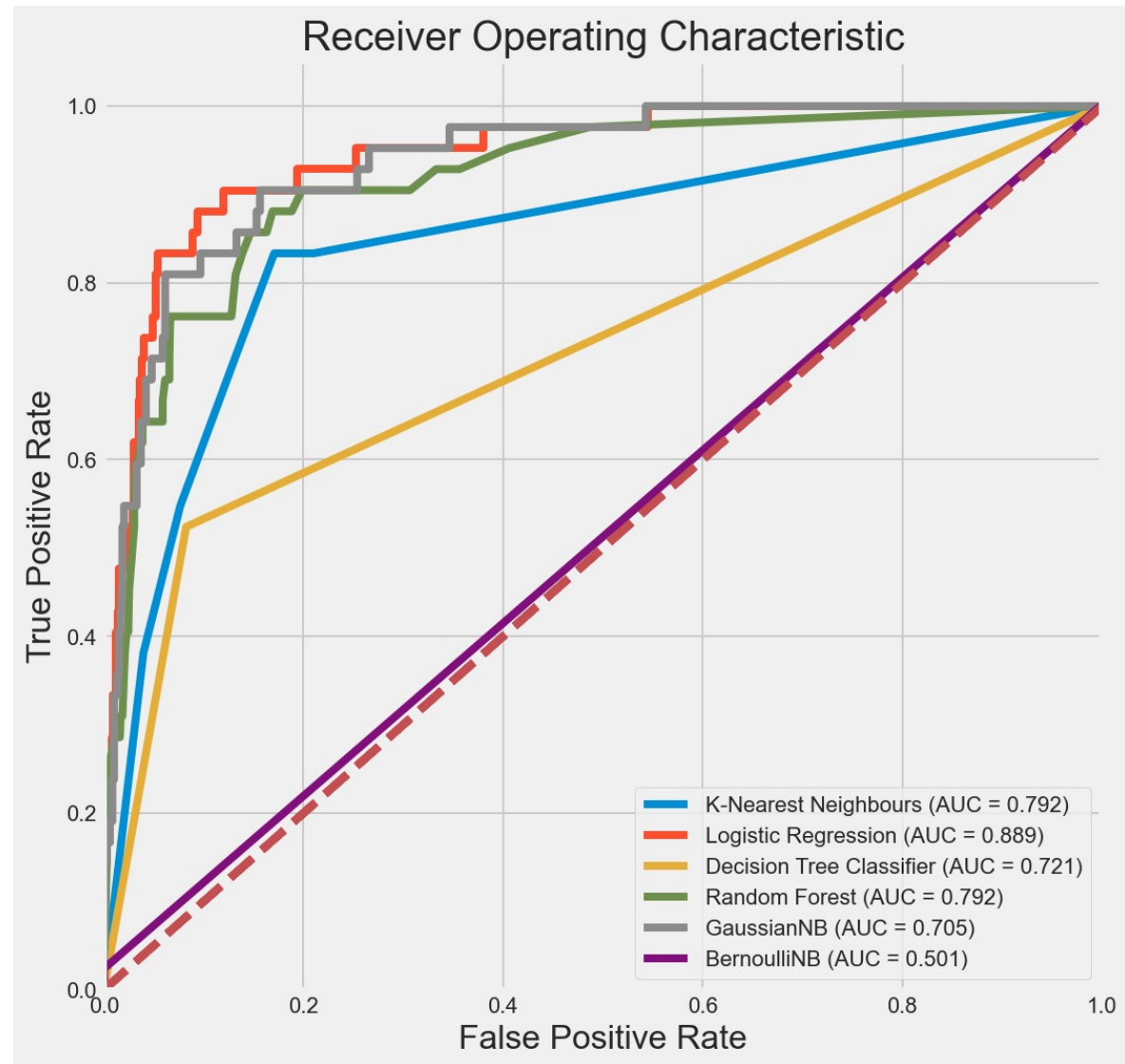- From 96 features to 31 features to corporate with S&P500 data

Receiver Operating Characteristic

Algorithm	Train Score	Model Score	Recall	Precision	F1 score	F-Beta	ROC-AUC score
Logistic Regression	88.86000	87.5%	0.88000	0.20000	0.32000	0.32000	0.88000
GaussianNB	82.36000	72.89%	0.95000	0.11000	0.19000	0.19000	0.84000
K-Nearest Neighbours	96.47000	89.94%	0.67000	0.20000	0.31000	0.31000	0.79000
Random Forest	100.00000	95.05%	0.55000	0.35000	0.43000	0.43000	0.76000
Decision Tree Classifier	100.00000	92.45%	0.45000	0.21000	0.29000	0.29000	0.70000
BernoulliNB	68.02000	39.94%	0.81000	0.04000	0.08000	0.08000	0.60000

Naïve Result

Top 2 Models (ROC-AUC & Recall)

- Logistic Regression

- Gaussian Naive Bayes

Tuned Result

- Removing multicollinearity

Receiver Operating Characteristic

- K-Nearest Neighbours (AUC = 0.792)
- Logistic Regression (AUC = 0.889)
- Decision Tree Classifier (AUC = 0.721)
- Random Forest (AUC = 0.792)
- GaussianNB (AUC = 0.705)
- BernoulliNB (AUC = 0.501)

Result

Logistic Regression confusion matrix:
- No Bankruptcy / No Bankruptcy: 1040
- No Bankruptcy / Bankruptcy: 150
- Bankruptcy / No Bankruptcy: 4
- Bankruptcy / Bankruptcy: 38
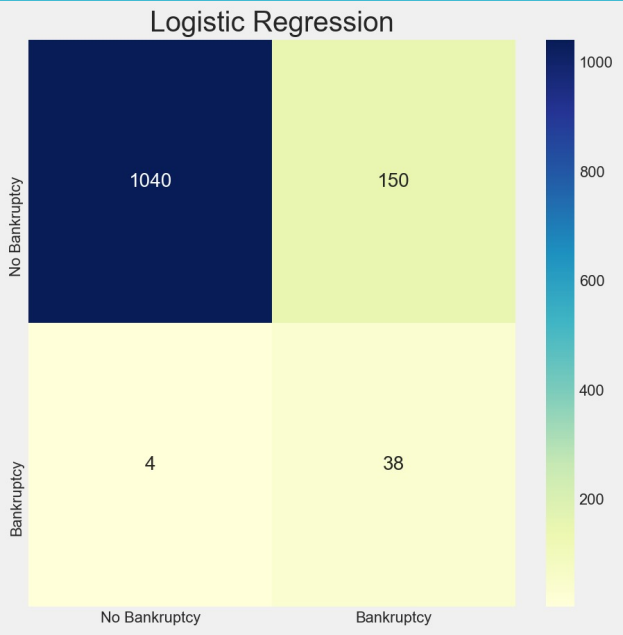
Model	Baseline	Tuned	
Recall	88%	90%	⬆ 0.02
Precision	20%	20%	
F1	0.32	0.33	⬆ 0.01
ROC_AUC	0.88	0.89	⬆ 0.01
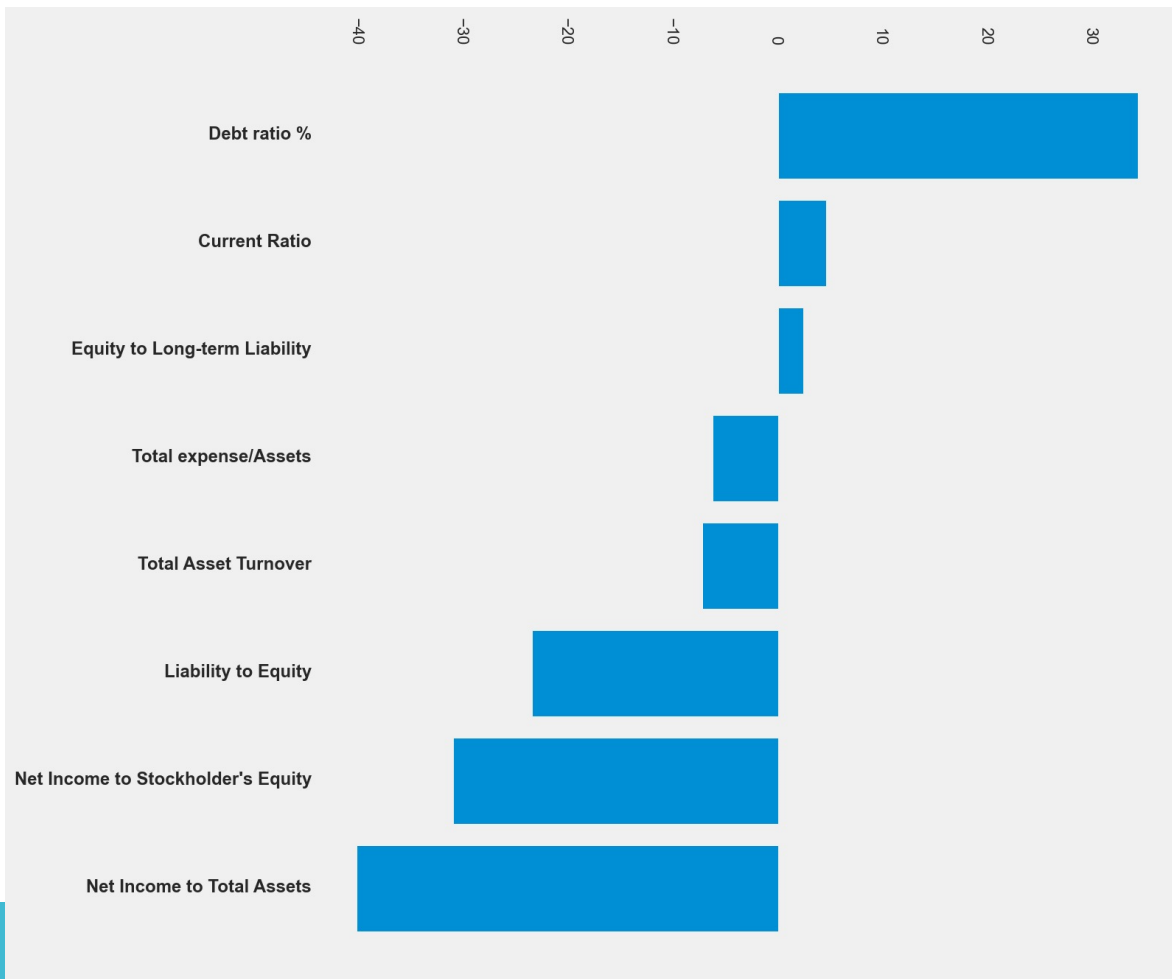
| Feature	Feature Coefficient
Debt ratio %	34.34437
Current Ratio	4.62239
Equity to Long-term Liability	2.40398
Total expense/Assets	-6.22252
Total Asset Turnover	-7.189
Liability to Equity	-23.468
Net Income to Stockholder's Equity	-30.9835
Net Income to Total Assets	-40.1931

Logistic Regression

S&P500 Top identified companies

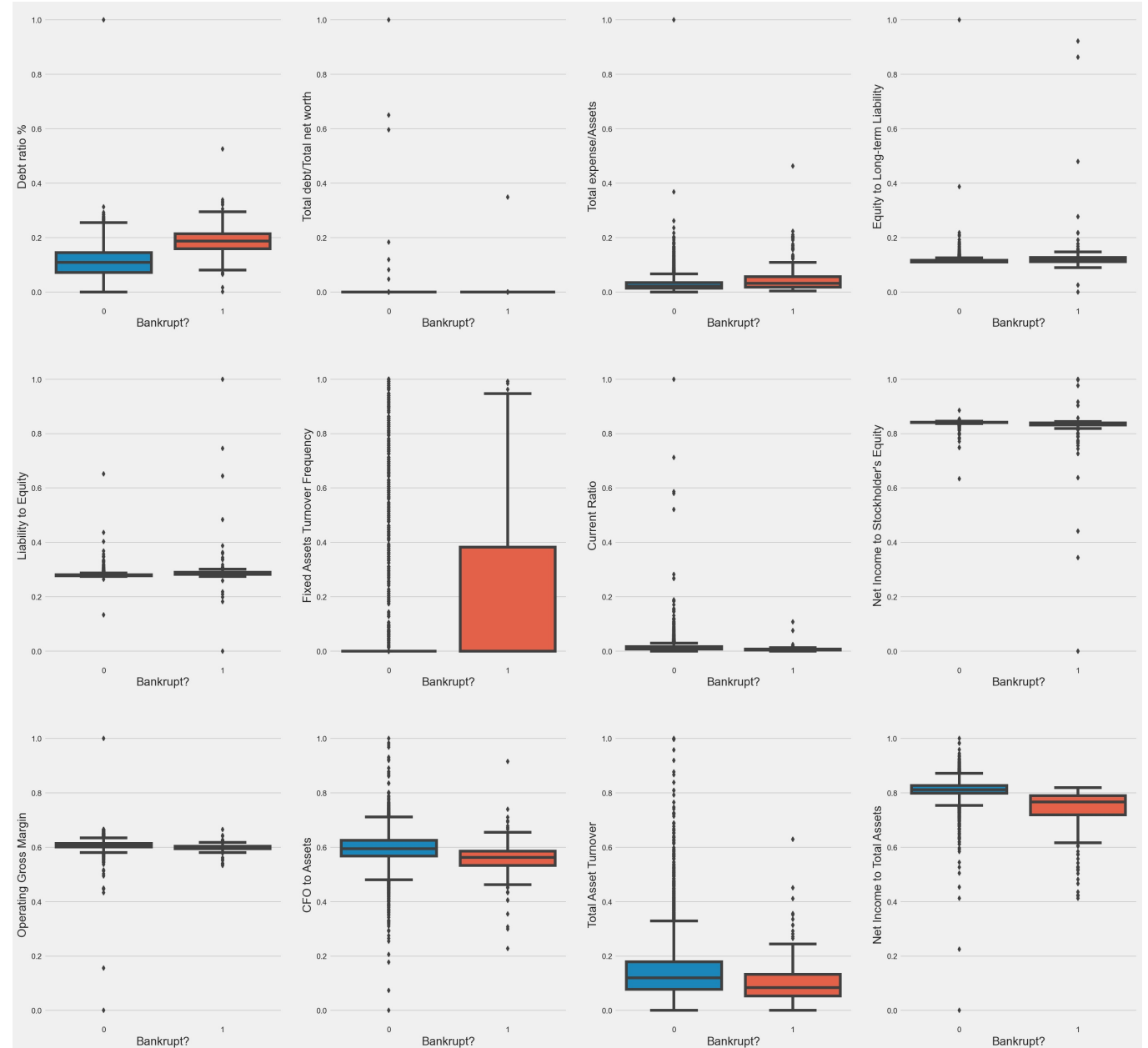|	Net Income to Stockholder's Equity	Debt Ratio%
TransDigm Group Inc	-0.32	1.2
MSCI Inc	-0.5	1.06
American Airlines Group Inc	-7.12	1.03
Kimberly-Clark Corp.	-34.58	1
Lamb Weston Holdings Inc	-180.65	1.02

Future works

- Try more sophisticated model techniques: XGBoost
- Try on Random forest (fix overfitting problem)
- Improving High False Positive
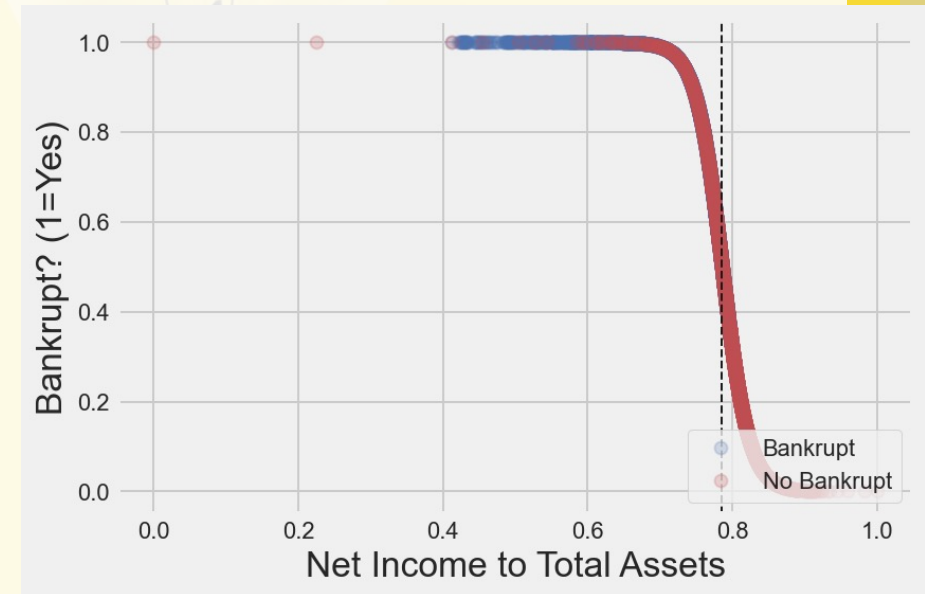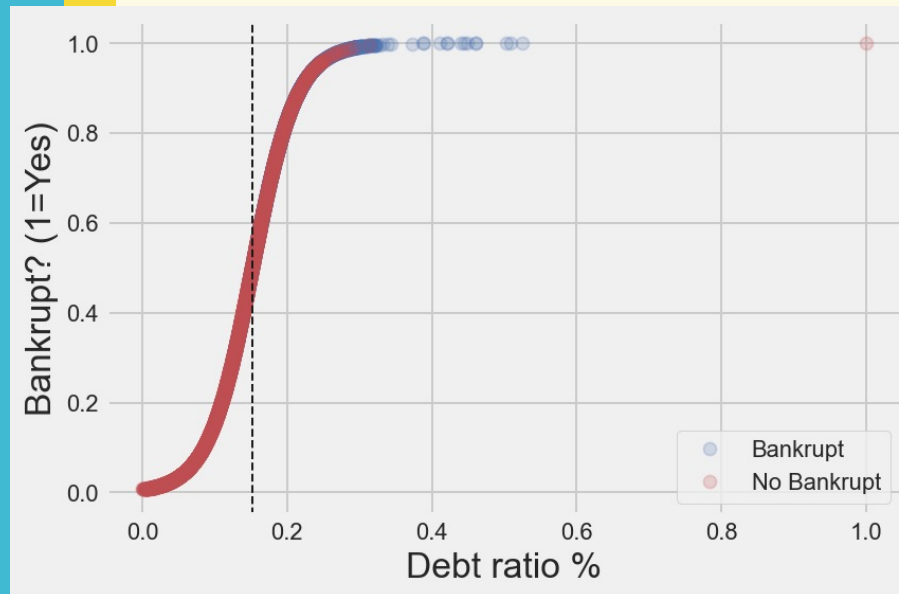
Thank you

Feature Classification Performance

Pairplot of 12 features