Predict Company Bankrupt

- Shijie Mao
- Brown University

- Dec. 7th . 2021
- GitHub Repo:

https://github.com/shijiemao/project1030

Recap

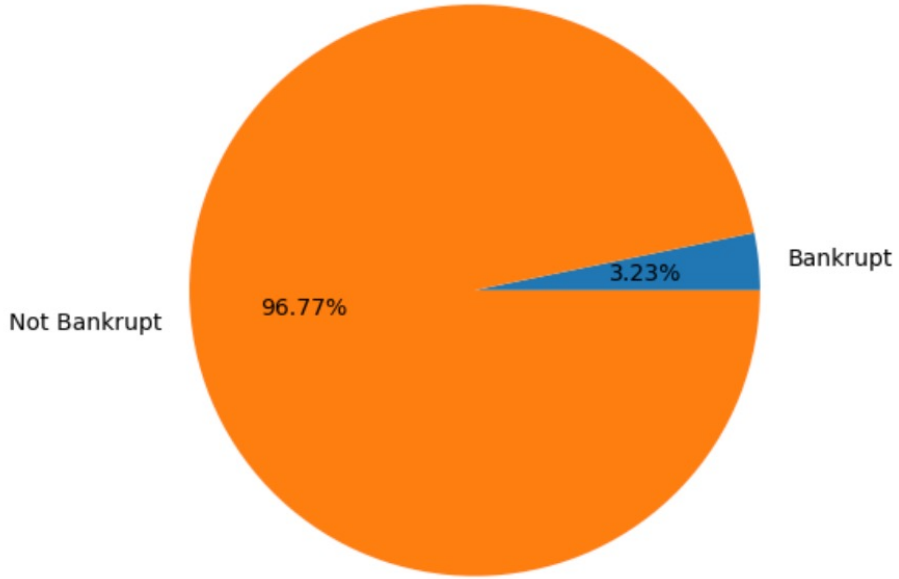
- Classification Problem
- Research on Bankrupt company which is essential for people who want to invest in these companies
- Build a model to predict the bankrupt company and find the most important features which influence the bankrupt.
- No missing values.
- One categorical variable only has one class
- Data source: Kaggle

https://www.kaggle.com/fedesoriano/company-bankruptcy-prediction

Cross Validation: Data Splitting

- Stratified K-fold: imbalanced data, no time-series pattern.
  only 3.2% of the target variable are class 1 (bankrupt).

- 20% of the data set as Testing data
- K = 4 to do cross-validation on the rest 80% of the dataset, using 20% data as Validation set each fold.
- Use the test data to calculate scores.



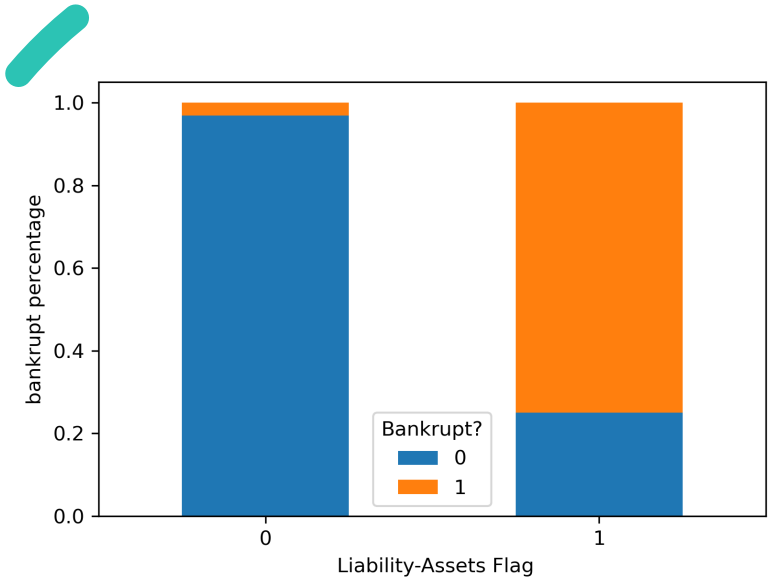
Cross Validation: Data Preprocessor

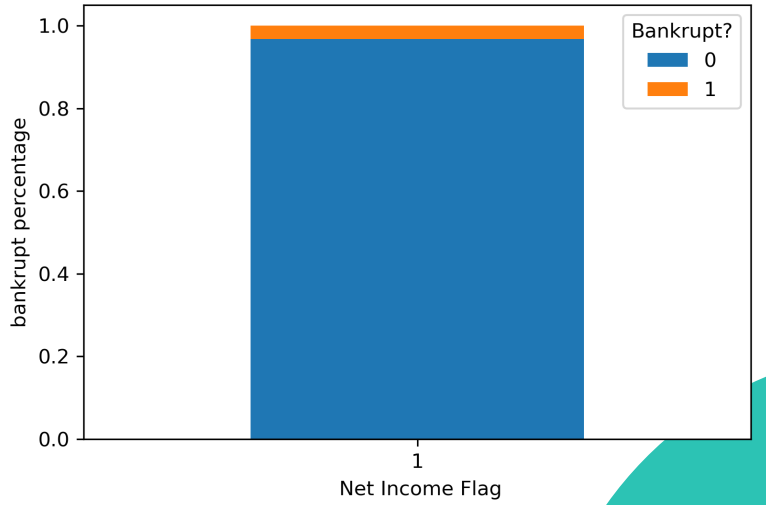
95 features: 93 numeric, 2 categorical

Standard Scaler for numeric

OneHot encoding for categorical

After preprocessing, 96 features expected.





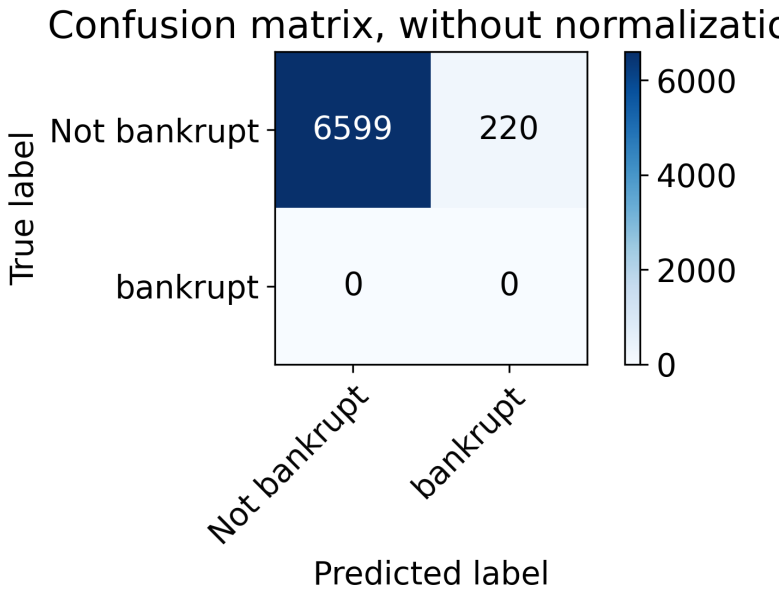
Cross Validation: Supervised ML Algorithms

Algorithms	Tuned Parameter	Parameter Value
Logistic Regression	C	np.logspace(-10,0,10)
Random Forest	Max_depth Max_features	[1,3,10,30,100] [0.2, 0.5, 0.75, 1.0]
Support Vector Machine Classification	Gamma C	[1e-2, 1e-1, 1e0, 1e1, 1e2] np.logspace(-1, 1, 5)
KNearestNeighbor Classification	N_Neighbors Weights	[1, 3, 10, 30, 50, 100] ['Uniform', 'distance']

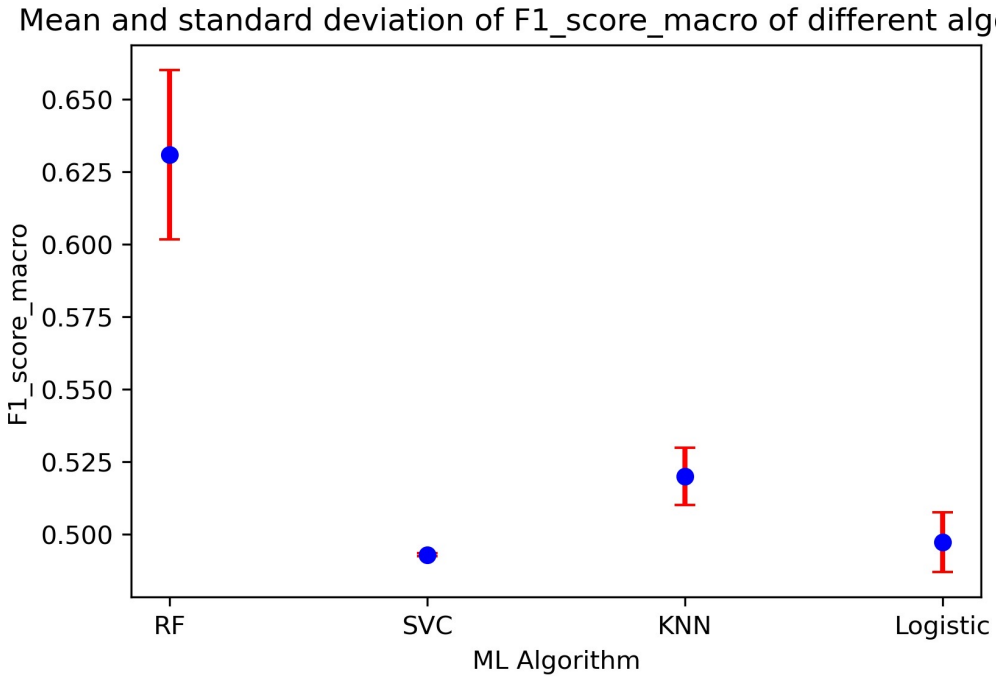
Result : Evaluation Metric and Baseline

- Using f1_score_macro as evaluation metric: Imbalanced data.
- Confusion Matrix of baseline situation (all 0, not bankrupt)

- F1_score_macro would be 0.49180205693844087



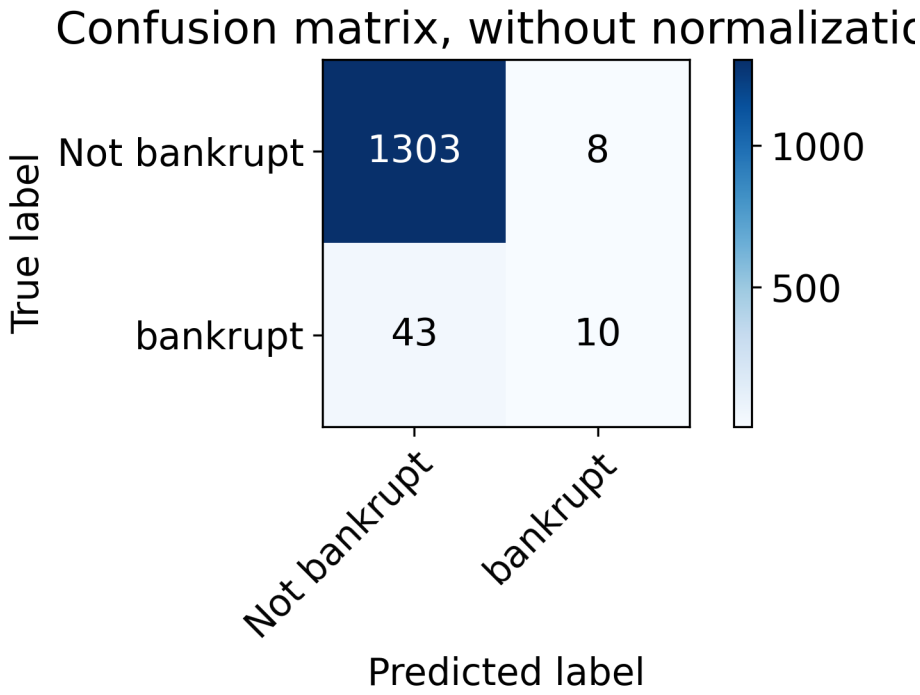
Result : Algorithms Performance (RF is the best)



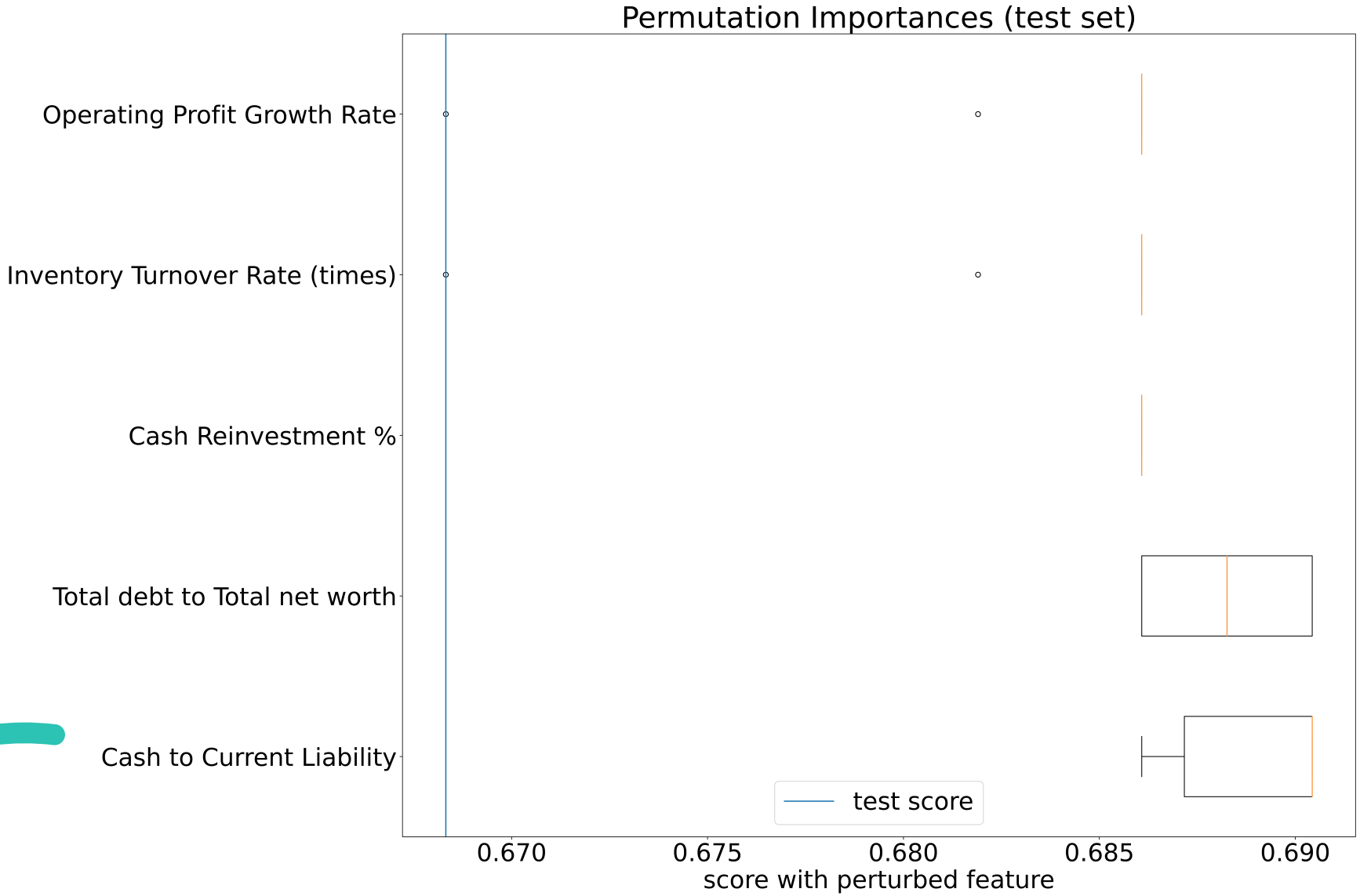
Random Forest with parameter max_depth = 30 and max_features = 0.2 has a highest scores with a mean of 0.63 and a standard deviation of 0.029. Which score is 4 standard deviations higher than the baseline score.

Result : Inspection in Random Forest

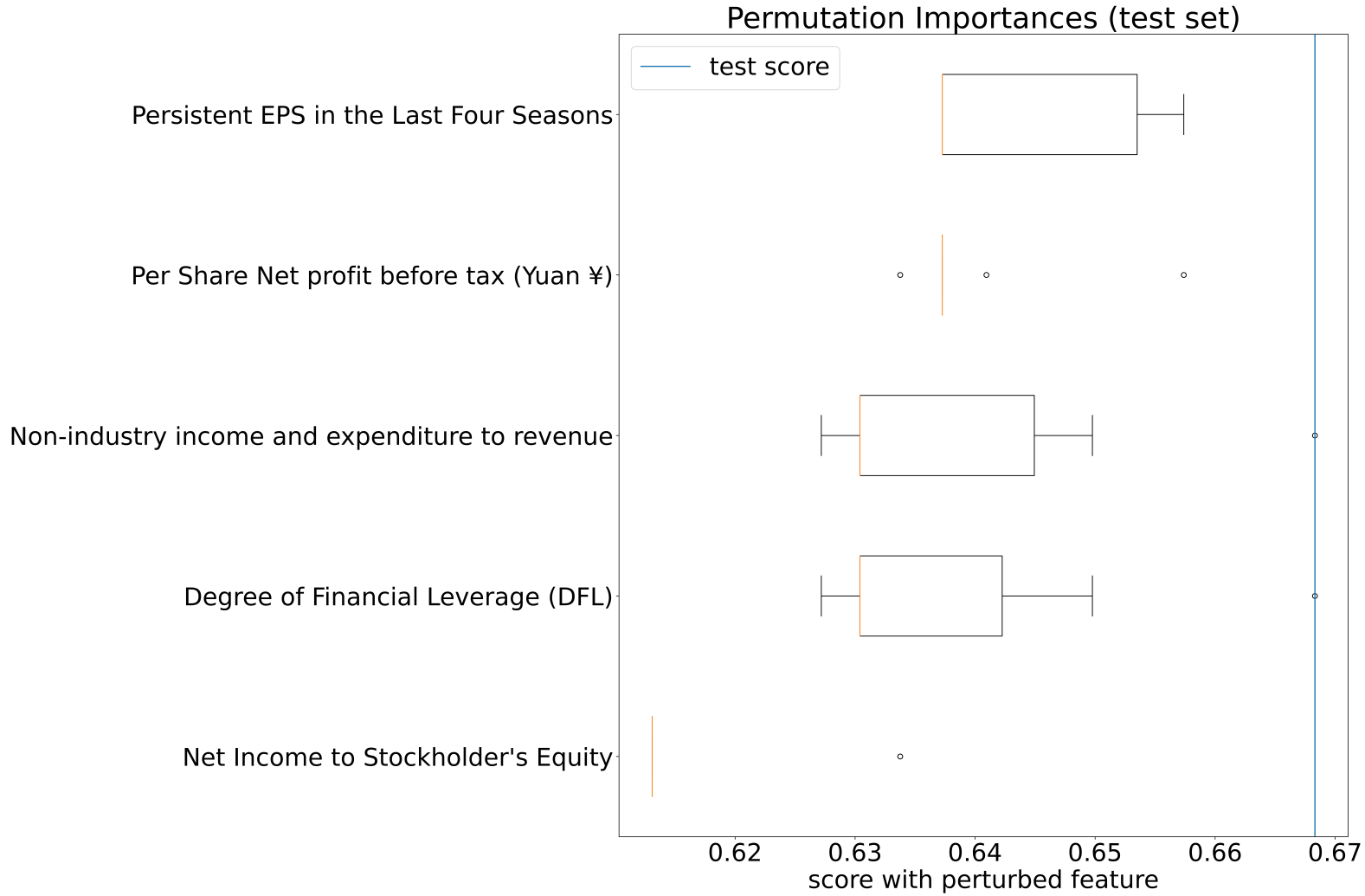
- Confusion Matrix
- Could identify half of bankrupt
- F1_score_macro
  = 0.631247780245644



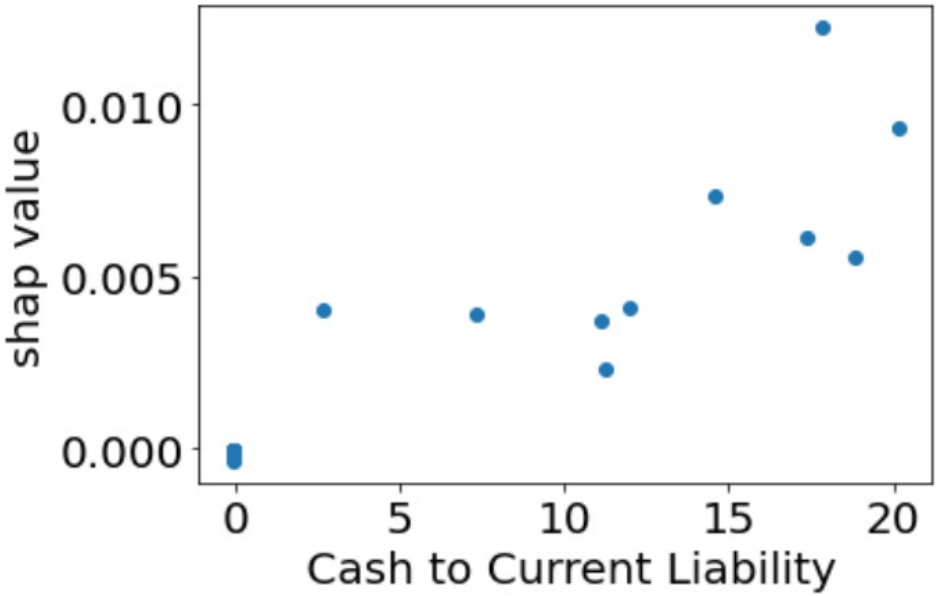
Result : Global Importance (5 most important)



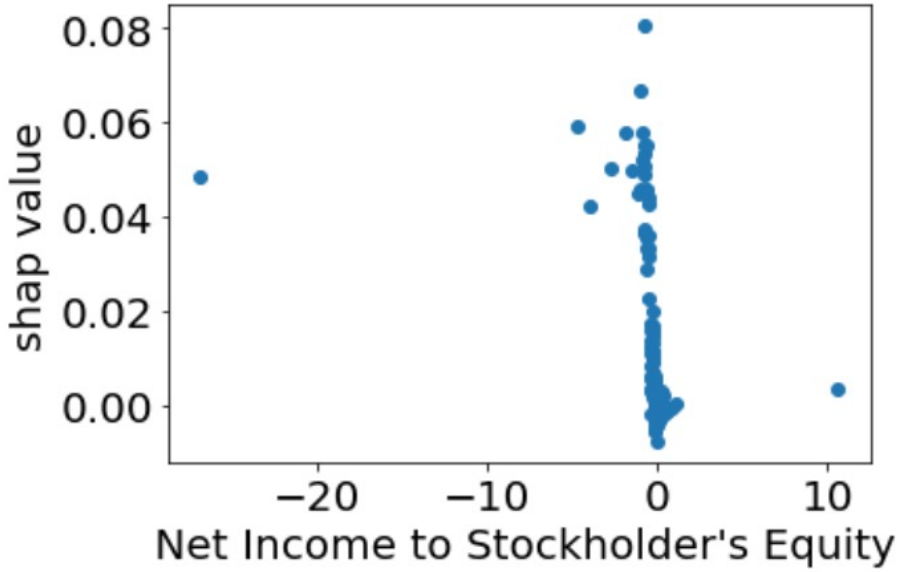
Result : Global importance (5 least important)



Result : Local importance of most & least feature



Most important



Least important

Outlook

- Tune parameter more precisely and diversity. Especially in SVC
- Think ways to keep most important Global feature stable.
- Interpret more local importance to give more meaningful insight of bankrupt company

Thank you!