Capstone Project

Company Bankruptcy Prediction

by
Ritik Vaidande
Data Science Practitioner
AlmaBetter, Bengaluru

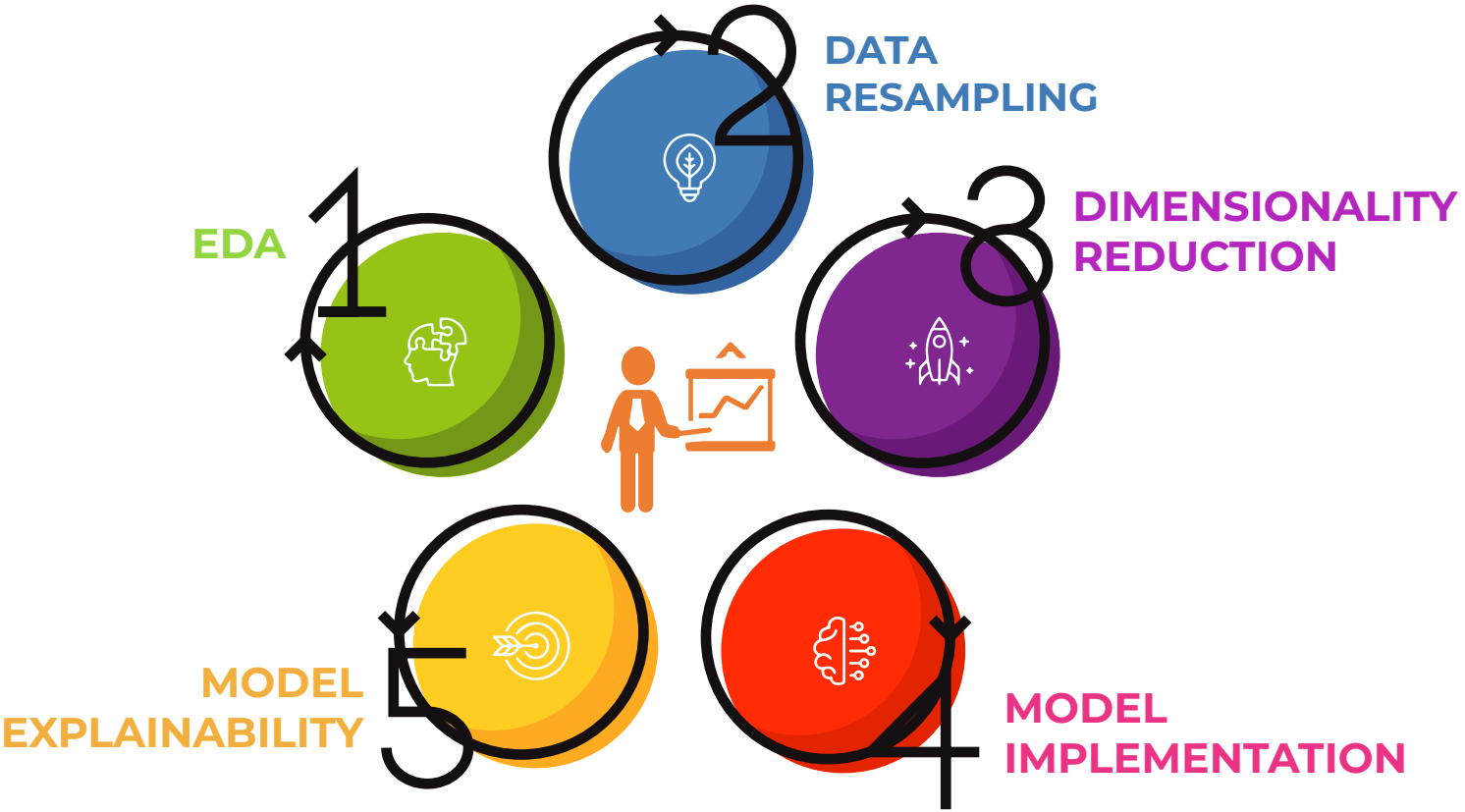
POINTS FOR DISCUSSION

1. Problem Statement
2. Model Architecture
3. Data Summary
4. EDA
5. Feature Correlation
6. Data Resampling
7. Model Implementation
8. Model Explainability

PROBLEM STATEMENT

Prediction of bankruptcy is a phenomenon of increasing interest to firms who stand to lose money because of unpaid debts. Since computers can store huge dataset pertaining to bankruptcy, making accurate predictions from them before hand is becoming important.

Predicting An Company Will Go Bankrupt Or Not


DATA
RESAMPLING
EDA
DIMENSIONALITY
REDUCTION
MODEL
EXPLAINABILITY
MODEL
IMPLEMENTATION
MODEL ARCHITECTURE

DATASET

Shape - (6819, 96)
Target Label -
0 : Not Bankrupt
1 : Bankrupt
Total Companies - 6819

The data were collected from the Taiwan Economic Journal for the years 1999 to 2009.

Current Liabilities / Equity

ROA(A) before interest and % after tax

Debt Ratio %

Current Liability To Current Assets

Current Liability To Assets

Borrowing Dependencies

Net Worth / Assets

ROA(B) before interest and depreciation after tax

Persistent EPS in the Last Four Sessions

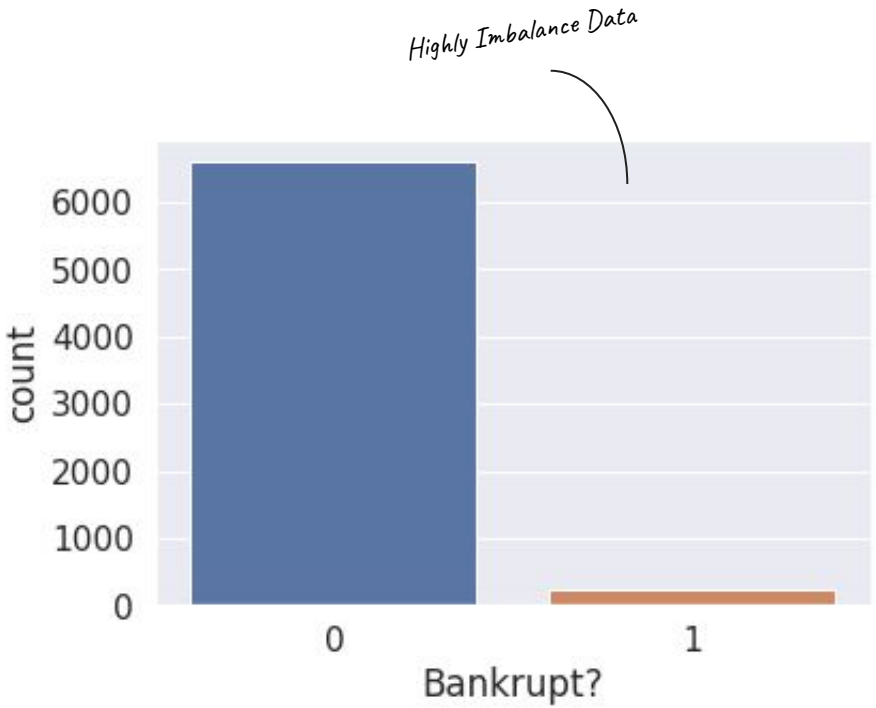
BANKRUPT COMPANIES

Non - Bankrupt Companies :

6599

Bankrupt Companies :

220



The target variable is highly imbalanced. Less than 3.5% instances for bankrupt companies exist in the data set

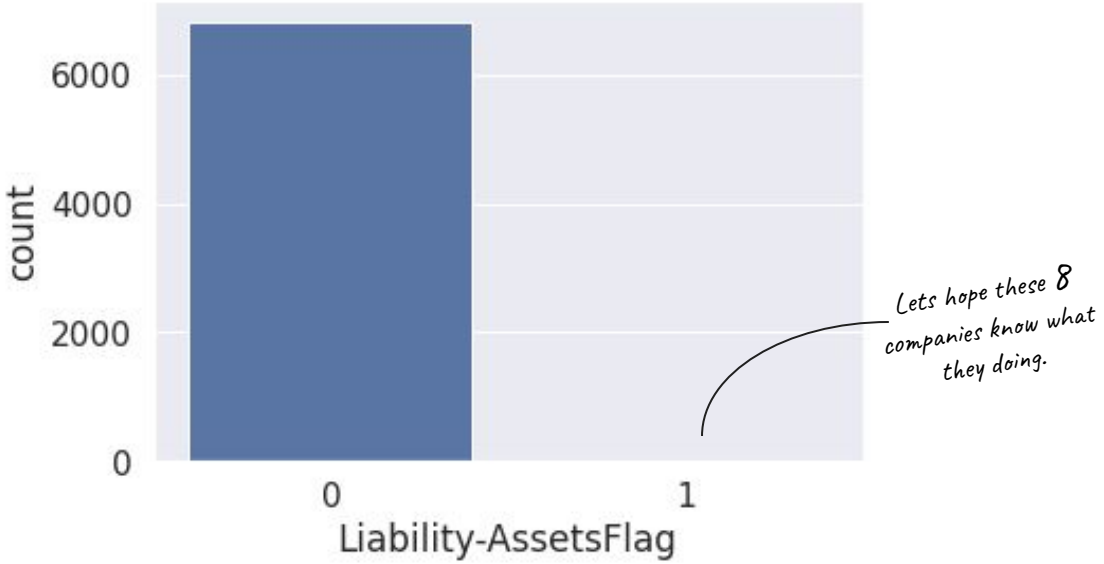
LIABILITY ASSETS FLAG

Companies Where Total Liability Exceeds Total Assets :

6811

Companies Where Total Liability is Less Than Total Assets:

8



The "Liability-Assets" flag denotes the status of an organization, where if the total liability exceeds total assets, the flagged value will be 1, else the value is 0.

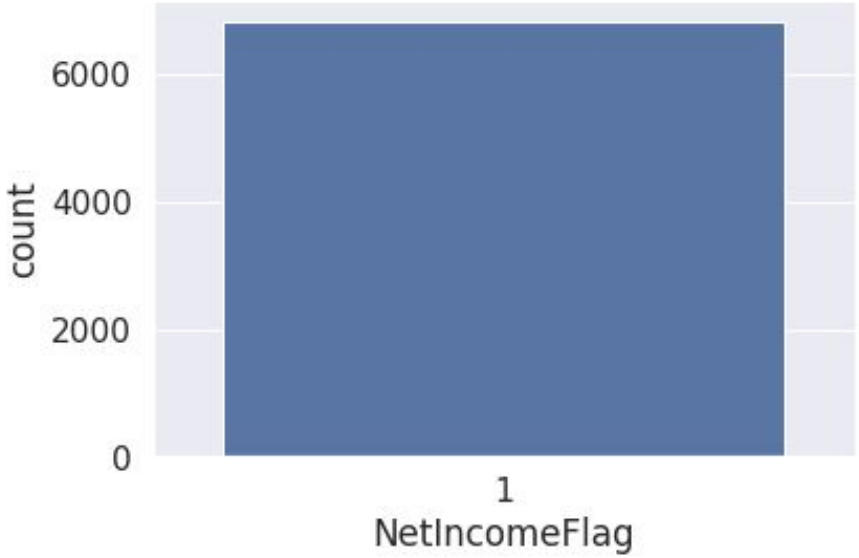
NET INCOME FLAG

Companies With Negative Net Income :

6819

Companies With Positive Net Income :

0



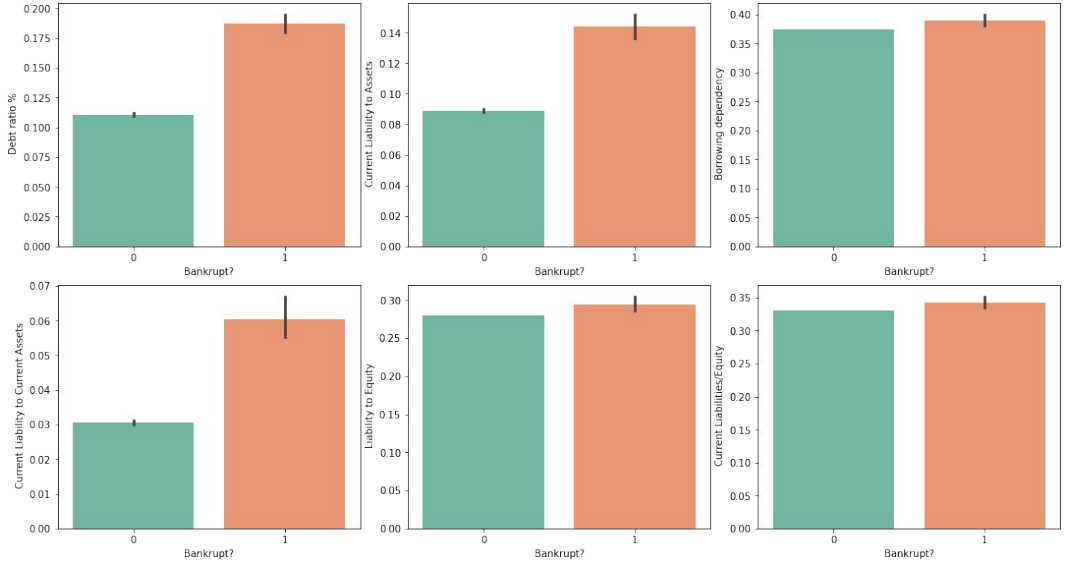
The "Net Income" flag denotes the status of an organization's income in the last two years, where if the net income is negative for the past two years, the flagged value will be 1, else the value is 0.

CORRELATION-MATRIX

There isn't much correlation between the features, most of the values lies around zero.

TOP 6 POSITIVELY CORRELATED ATTRIBUTES

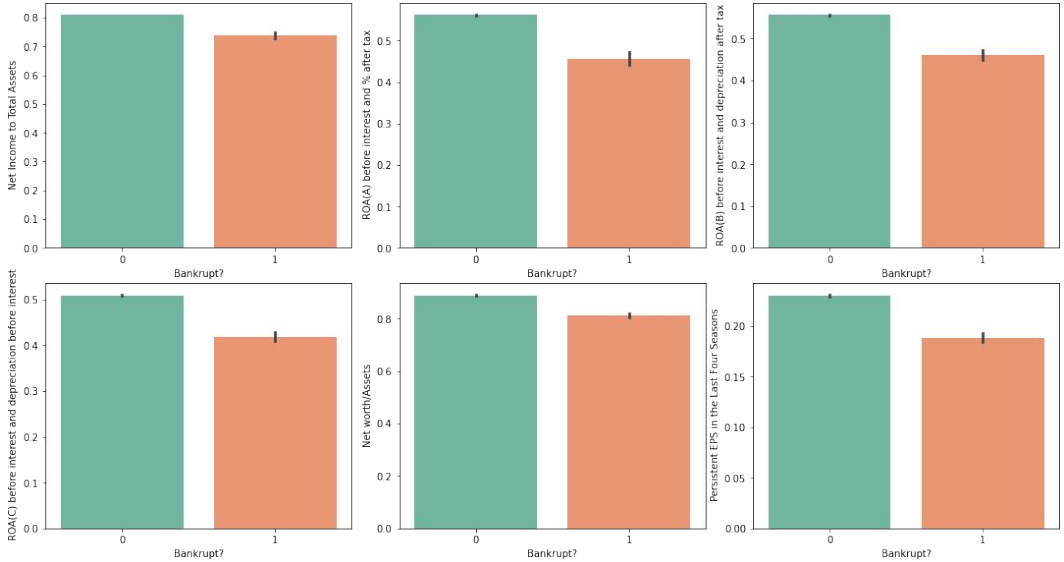
1. Debt Ratio %
2. Current Liability to Assets
3. Borrowing Dependency
4. Current Liability to Current Assets
5. Liability to Equity
6. Current Liabilities/Equity



In bankrupt organizations attributes such as Debt Ratio %, Current Liability To Assets, Current Liability To Current Assets are found to be higher.

TOP 6 NEGATIVELY CORRELATED ATTRIBUTES

1. Net Income to Total Assets
2. ROA(A) before interest and % after tax
3. ROA(B) before interest and depreciation after tax
4. ROA(C) before interest an depreciation before interest
5. Net Worth/Assets
6. Persistent EPS in the Last Four Sessions



These plot shows that an organization is less likely to go bankrupt when they typically earn more and hold more assets.

SMOTE

Original Training Target Variable :

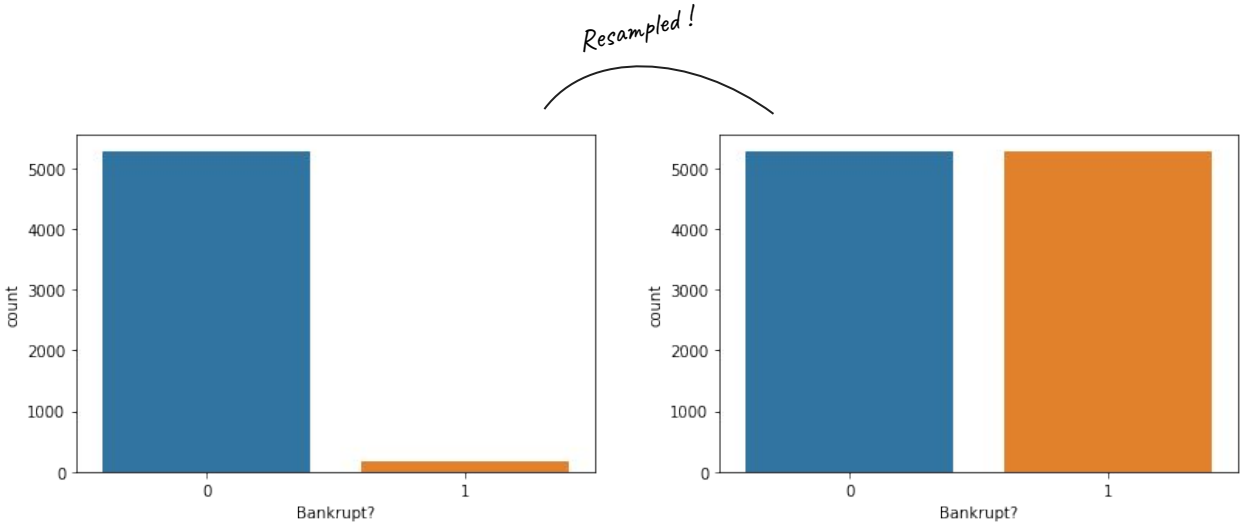
0 : 5286

1 : 169

Training Target Variable after SMOTE :

0 : 5286

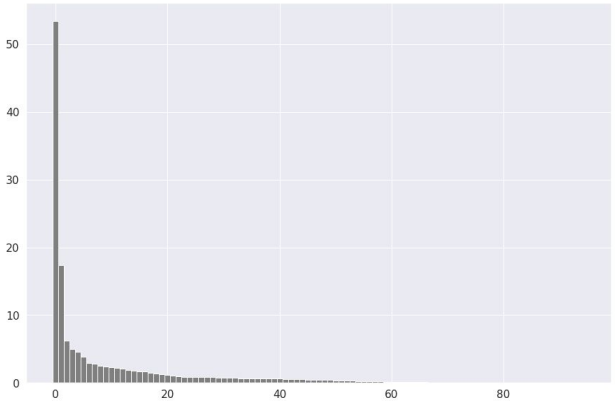
1 : 5286

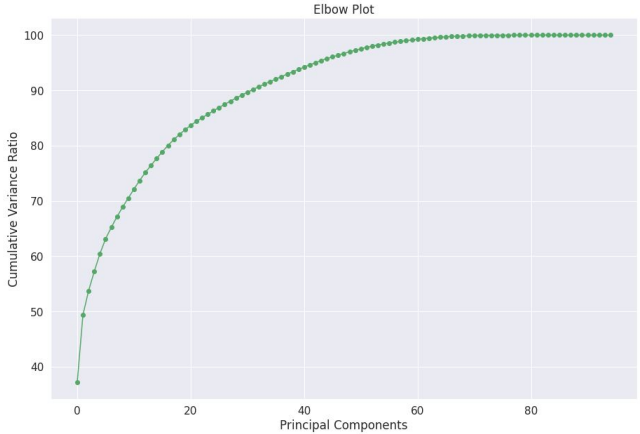


Synthetic Minority Oversampling Technique (SMOTE) is a statistical technique for increasing the number of cases in your dataset in a balanced way. The component works by generating new instances from existing minority cases that you supply as input.

PCA

Principal Component analysis, or PCA, is a dimensionality-reduction method that is often used to reduce the dimensionality of large data sets, by transforming a large set of variables into a smaller one that still contains most of the information in the large set.





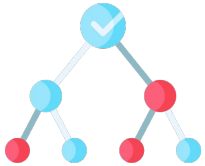
100% of the variance could be explained by almost 60 components.

MACHINE LEARNING CLASSIFICATION MODELS IMPLEMENTED

Random Forest Classifier

XGBoost XGBoost

Logistic Regression



Decision Trees Classifier

Support Vector Classifier

MODEL PERFORMANCE

Best Performing Model :

XGBoost

Least Performing Model :

Random Forest Classifier

best_score is calculated as an average of the cross-validation iterations.

	model	best_score	best_params
4	XGBoost	0.968976	{'XGBoost__n_estimators': 400}
2	DecisionTC	0.956584	{'DecisionTC__min_samples_split': 2}
1	SVC	0.947219	{'SVC__C': 2}
3	RandomFC	0.913734	{'RandomFC__max_depth': 4, 'RandomFC__min_samp...
0	Logistic	0.903992	{'Logistic__C': 2, 'Logistic__solver': 'newton...

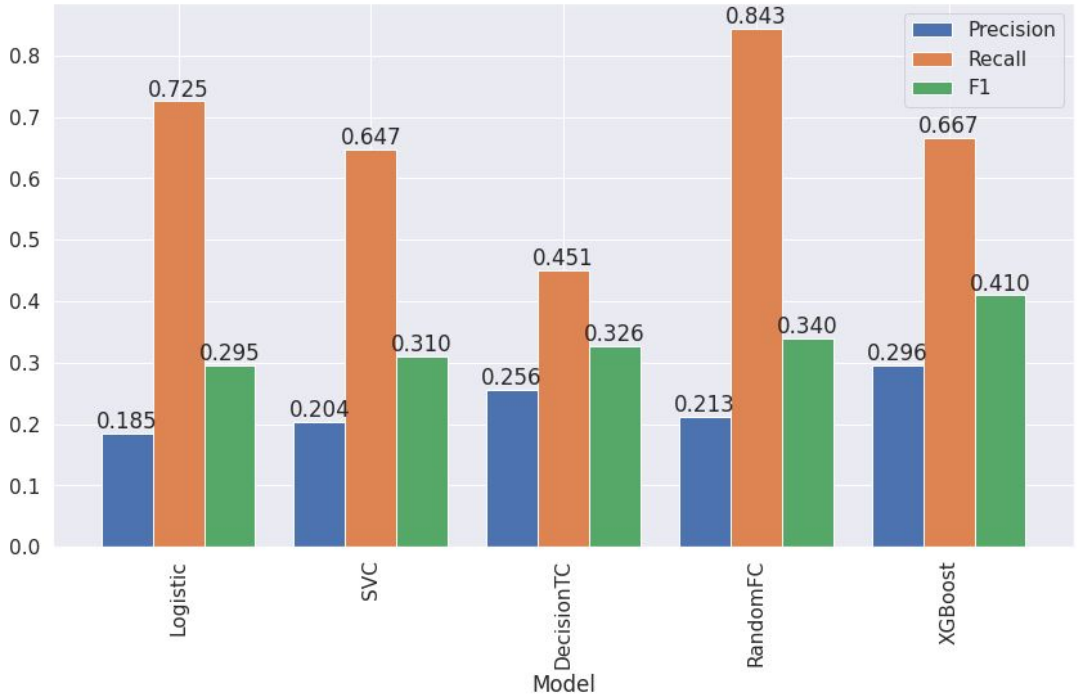
XGBoost and Decision Trees Classifier achieved good results while Logistic Regression and Random Forest Classifier gave satisfactory performance.

MODEL PERFORMANCE

Highest Precision :
XGBoost

Highest Recall :
Random Forest Classifier

Highest F1 Score :
XGBoost



MODEL SELECTION

Model Selected :

XGBoost

Best Hyperparameter :

XGBoost__n_estimators = 500

XGBoost is selected for the classification problem by considering the highest F1 score.

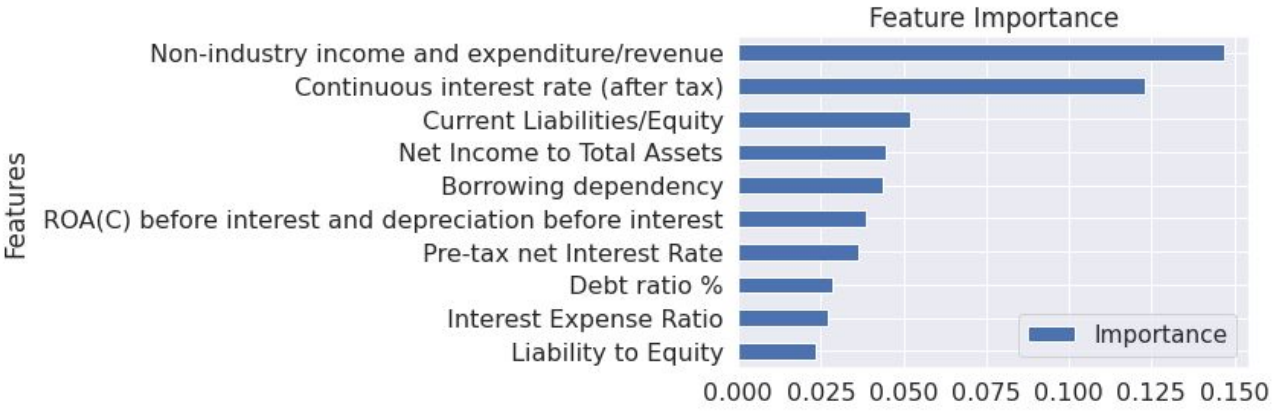
FEATURE IMPORTANCE

Model :

XGBoost

Most Important Feature :

Non-industry income and expenditure/ revenue



These are the top 10 most important features as per the top performing model XGBoost.

CONCLUSION

There are many attributes that play important roles to decide whether a company will go bankrupt or not.

Net Income Flag plot showed us that most of the companies are running into Losses for the past 2 years.

There are high chances that a company can go Bankrupt if the attributes “Debt Ratio %, Current Liability To Assets, Current Liability To Current Assets” are high.

The best performing model is XGBoost by considering the F1 score which is an ideal metric to choose for an classification model.

Analyzing the dataset and building the best model to predict bankruptcy is been done successfully.

Thank You

This presentation is an part of the Classification Capstone Project by Ritik Vaidande, AlmaBetter.