**Objective**

In this case study, the team will build a predictive model for determining what factors could indicate that a company will go bankrupt. The team will review the historical data and determine the prime indicator that will allow the investor to divest the company ahead of time.

**1 Introduction**

Maximizing profits is the prime pursuit of every business. On another side, the business must have the ability to generate enough revenue and margins to pay debts. The companies' inability to pay off their liabilities in a timely can lead to Bankruptcy. "Bankruptcy is a legal proceeding involving a person or business that is unable to repay their outstanding debts" [1]. Bankruptcy is the culmination of multiple issues leading to the economic collapse of a company certainly warrants detailed analysis to identify the early warning indicators to trigger mitigation sooner.

Bankruptcy is a very complex subject. The factors leading to Bankruptcy may vary from one business type to other. So, it is critical that matrices need to be evaluated from multiple dimensions to ensure the randomness of the cause gets the appropriate attention. The team decided to look at a classification model that looks at the challenge with the different permutation and combination of input factors and create an ensemble of different outcomes to generate the final outcome. The best method to execute this methodology team decided to go with Random Forest. The Random Forest Algorithm generates multiple individual trees from attribute selection and independent random samples. The outcomes of each tree are averaged to determine the end result.

Another critical factor was finding the right balance between variance and bias. Similar ensemble learning technique XGBoost (Extreme Gradient Boosting) will be leveraged to build a robust classifier with the larger categorical and numerical dataset. XG boost classifiers work very similar to Random Forest. This classifier built multiple trees like Random foster. Rather than averaging the outcomes of trees, it feeds to the outcome from one tree to another to enhance the model performance.

This case study aims to build a classification model that predicts whether a company will enter Bankruptcy based on a wide array of collected finical information about those companies. Companies still in operation came from 2007 to 2013. The authors built an Extreme Gradient Boosting model using "synthetic features" they created from the accumulated data, which they described as "a combination of the econometric measures using arithmetic operations (addition, subtraction, multiplication, division)" [2]. These synthetic features were used as the predictors in their model, and this case study deploys similar methods using the same synthetic features they derived. The authors analyzed each year in a five-year forecasting period, but this case study only collectively models the cumulative data from all five years.

**2 Methods**

#### Data Acquisition and Cleanup (Data Wrangling)

The data used in this case study is from the UCI Machine Learning database website [3]. The dataset in its original form is provided as five comma-separated ".arff" files for each of the five years of the forecast period; these were first combined in a single data frame for this cumulative analysis. There are no names, keys, or indicators of any kind for the included instances that each represents an evaluated company, and the data is purely based on economic indicators with no references to company sizes, types, industries, or vintages. All the variables were contiguous. The 64 column names in the dataset use alphanumeric keys (from X1 to X64) instead of the full description of the synthetic feature for simplicity and ease of use with coding. The full names of each synthetic feature are included in the Appendix below. The target class is Y, representing a bankruptcy with "1"

Based on visual inspection of the missing values, there appeared to be very few features with large amounts of missing data (Attr21 and Attr37). All missing values for each feature were replaced with the median of the existing data for that feature. The mean was not used because some features have extreme outliers that could influence the mean (refer to Box lots below). It was also determined that no other features could be used with significant benefit to predictively impute the missing values in the few features with a high number of missing data**.**

The dataset has various metrics that are of different magnitudes. The tree-based ensemble methods can handle this data effectively. However, the team decided to normalize the data.

Note: All imputation and normalization were carried out post splitting of data into training and test data independent of each other to prevent data leakage between training and test data set.

**Metrics**

The metrics used to evaluate model performance were accuracy and False Negative. So, the team had to find the right balance between accuracy and False Negative. The Recall will be the prime focus as Recall is penalized by False Negative.

Accuracy

Accuracy refers to the level of agreement between the actual measurement and the predicted value.

Recall:

These metrics are defined as follows:

Recall = TP/(TP+FN) = (Bankruptcy Correctly Identified) / ((Bankruptcy Correctly Identified) + Bankruptcy Incorrectly labeled as not Bankruptcy)

Where:

True Positive (TP) is "Bankruptcy is correctly predicted to be Bankruptcy". These predictions will correctly identify spam mail.

False Negative (FN) is "Bankruptcy is incorrectly predicting Solvency to be Bankruptcy". These predictions would predict ham as spam.

False Positive (FP) is "Bankruptcy is incorrectly predicting ham to be Bankruptcy". These predictions would identify ham as spam.

In the absence of stakeholder guidance, the team chose this balanced approach to misclassification.

**Modeling Objective**

The team's objective was to maximize accuracy and minimizing the Recall.

**Data Imbalance**

When we look at the distribution of the target outcome (i.e., whether a company went bankrupt or not), we see that the data is heavily imbalanced. As we would expect, the proportion of companies that went bankrupt is significantly less compared to the ones that did not.



Team decided to use stratified splitting technique during the train-test split process to ensure the proportion of target variable remains same in both the train and test data sets.

The purpose of this modeling exercise is to predict companies that may go bankrupt so that the business can make relevant divestiture decisions or trigger other mitigation strategies to minimize the losses. The effort is to minimize false negatives (i.e., misclassifying the companies that went bankrupt), so we have used the recall metric to compare the models during hyperparameter tuning, as the recall is the percentage of total positively identified bankruptcies correctly classified by the algorithm.

**Random Forest Model**

The dataset was first split into training and test sets with 30% of the data reserved for the test set. The target class is heavily imbalanced as discussed above, with only about 5% of the outcomes being bankruptcy. So, the test/train split was accomplished using the *stratified* option to maintain the ratio of outcomes.

Hyperparameter Tuning??

The Random Forest Model was tuned using GridSearch to iterate through different possible hyperparameter combinations. The primary metrics for Training the model and evaluating the hyperparameter combination was the recall on the test data set.

* *n\_estimators*: The number of trees in the forest.
* *max\_features*: The number of features to consider when looking for the best split.
* *criterion*: Gini or Entropy; The function used to evaluate the quality of a split.
* *max\_depth*: The maximum depth of the tree.

**Boosted Model**

**3 Results**

**4 Conclusion:**

**Final Model Proposal**

**Future Considerations and Model Enhancements**

Reference :

[1] [Bankruptcy Definition (investopedia.com)](https://www.investopedia.com/terms/b/bankruptcy.asp#:~:text=1%20Bankruptcy%20is%20a%20legal%20proceeding%20carried%20out,chapter%20within%20the%20U.S.%20...%20More%20items...%20)

[2] [Ensemble boosted trees with synthetic features generation in application to bankruptcy prediction - ScienceDirect](https://www.sciencedirect.com/science/article/abs/pii/S0957417416301592?via%3Dihub)

[3] [UCI Machine Learning Repository: Polish companies bankruptcy data Data Set](https://archive.ics.uci.edu/ml/datasets/Polish+companies+bankruptcy+data)