**PROJECT REPORT ON**

**COMPANY BANKRUPTCY PREDICTION MODEL**

**DONE BY:**

SHANMUKH VADDADI

**Introduction:**

In this project we will use various predictive models to see how accurate they are in detecting whether we can correctly predict which companies will face bankruptcy in the future.

**Overview of the Dataset:**

We had obtained the dataset from the Kaggle Public Dataset

Link: [Company Bankruptcy Prediction | Kaggle](https://www.kaggle.com/datasets/fedesoriano/company-bankruptcy-prediction)

This Dataset contains the information collected from the Taiwan Economic Journal for the years 1999 to 2009. Corporate bankruptcy is defined based on the business regulations of the Taiwan Stock Exchange. This dataset contains 6819 rows × 96 columns :

* 95 features (X1-X95, business regulations from the Taiwan Stock Exchange)
* 1 vector label

**Objectives to be achieved:**

**Main Objective:**  Predict the probability of a company going bankrupt based on the data provided.

**Goals:**

* Understand the key factors that influence a company's bankruptcy risk.
* Optimize the prediction model through hyperparameter tuning to get better accuracy.
* Deploy optimized models in production environments so they can be used for real-time analysis.

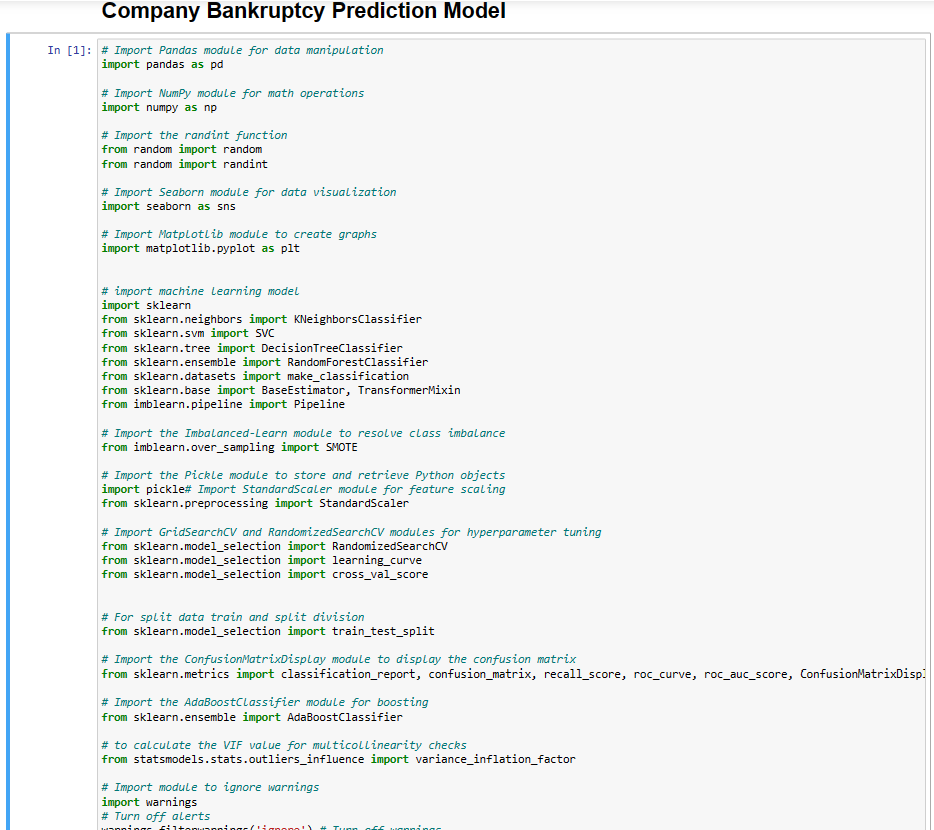
**Problem Statement:**

In an era of increasingly sophisticated stock trading, the ability of investment companies to understand and predict the risk of bankruptcy of a company is not only a competitive advantage but also a necessity. for an Investment Company, understands the importance of digging this information in depth. Protecting our clients' investments, as well as ensuring creditors, employees and management of the companies we invest in have a clear picture of their financial health is a top business priority.

**Contents in the Project:**

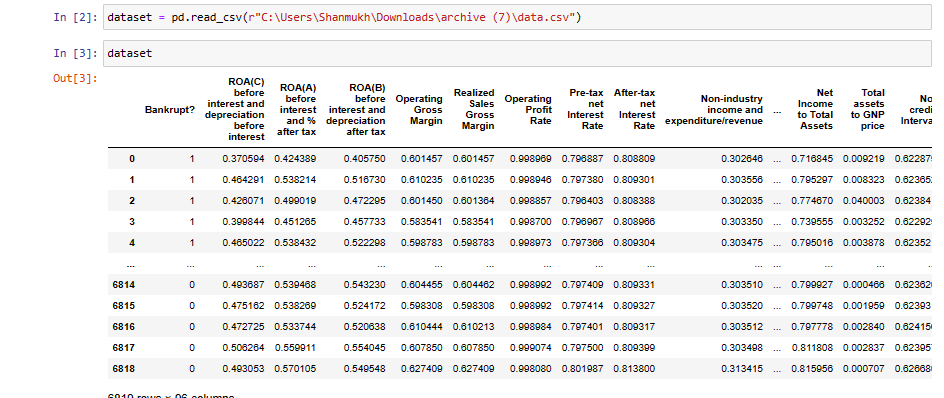
* Importing Libraries
* Data Loading
* Exploratory Data Analysis(EDA)
* Model Definition
* Model Training
* Model Saving
* Conclusion

**Importing Libraries:**

****

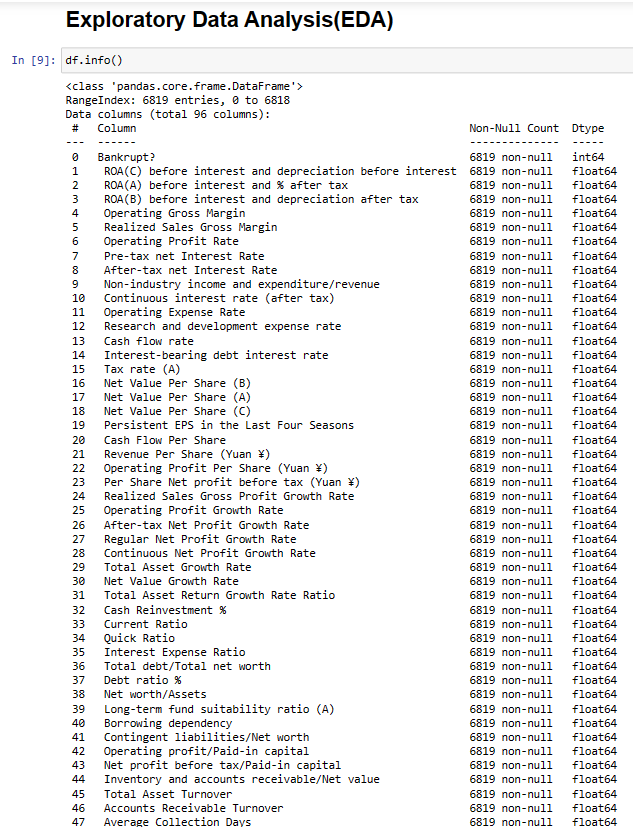
Initially we had the imported the libraries pandas, numpy, seaborn, matplotlib, sklearn, scipy, sckitlearn etc for the data cleaning and data analysis process, Data Visualization, Linear Algebra, Model Training, Model Evaluation, Confusion Matrix etc

**Dataset Loading:**

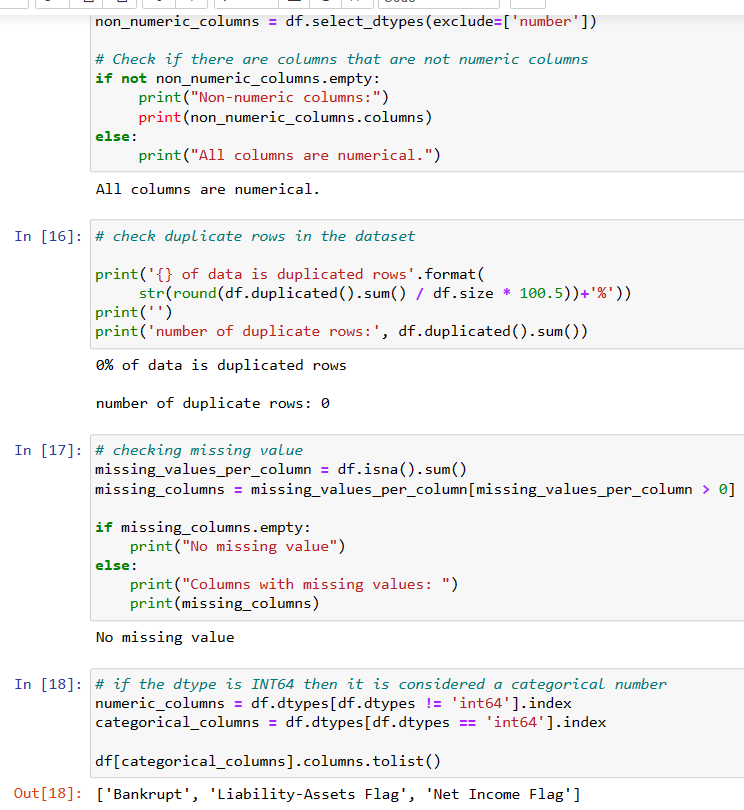
****

We had read the dataset file using pandas library and named it as ‘dataset’.

**Exploratory Data Analysis:**

****

The data content consists of 6819 rows and 96 columns.



All columns are numerical in the dataset.

There are no duplicate rows and missing values.

Looking from 'data.info()' that dataset has mostly "float64" data.

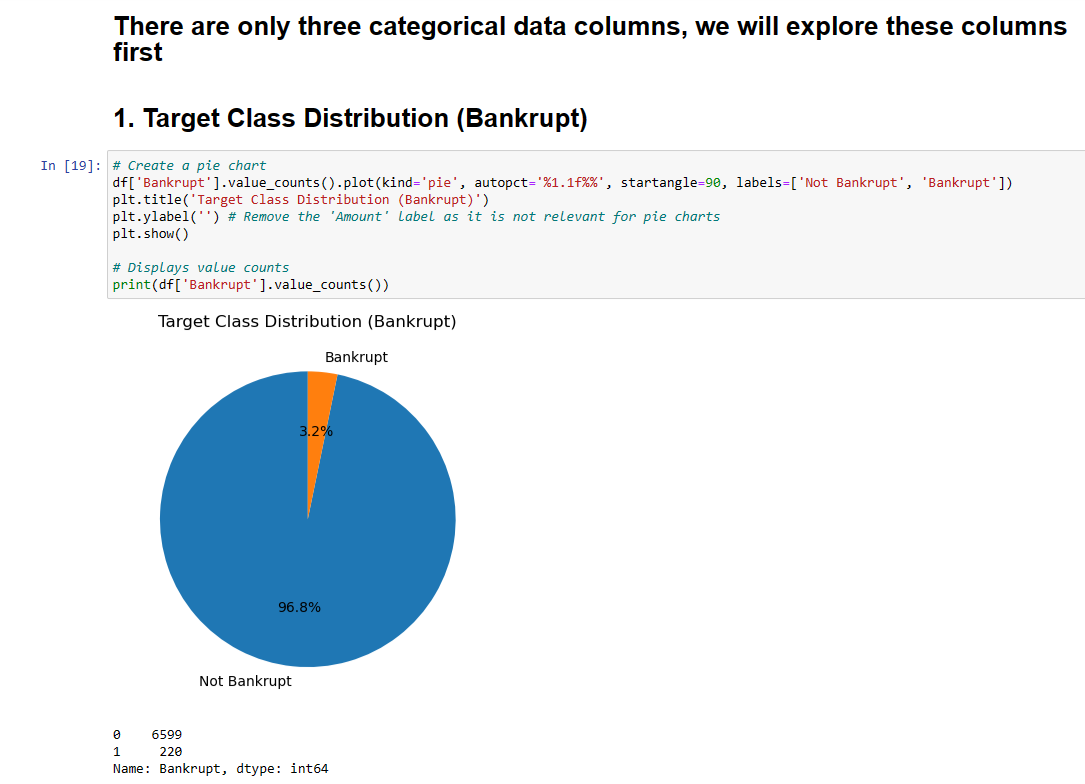
Categorical data is differentiated as binary numbers 1 and 0, so it is stored as type "int64". So we will separate numerical and categorical data to analyze the data set.

In this dataset we have three categorical

columns they are {‘Bankrupt’, ’Liability-Assets Flag’, ‘Net Income

Flag’} remaining are the numerical columns.

**Target Class Distribution**

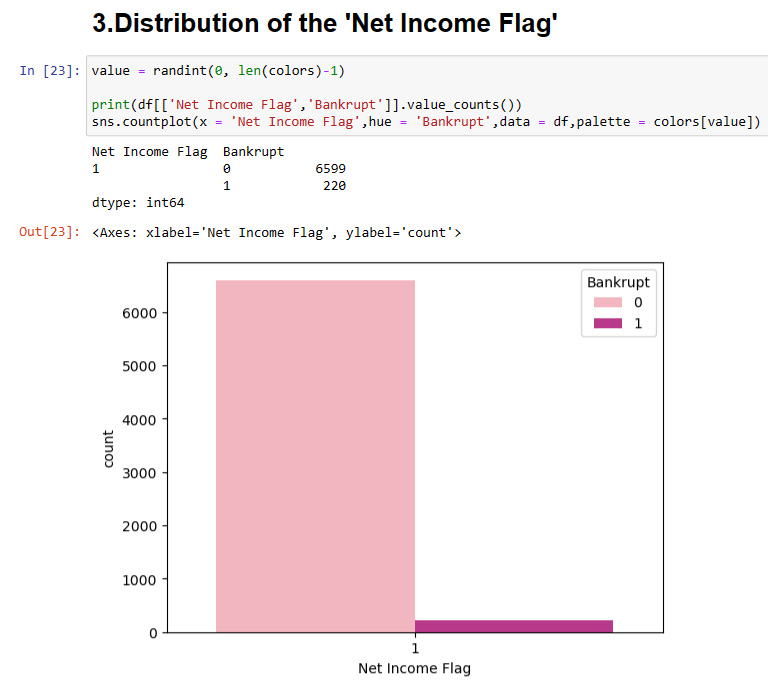
****

The data set looks very unbalanced. Therefore, it is necessary to consider balancing the dataset via the SMOTE technique.

SMOTE (Synthetic Minority Over-sampling Technique) is a technique used to overcome the problem of dataset imbalance. This is a technique commonly used to improve minority classes in classification problems.



The "Liability-Assets Flag" shows the status of an organization, where if total liabilities exceed total assets, the marked value is 1, otherwise the value is 0. Often, an organization/company's assets are greater than its liabilities. A small percentage of organizations go bankrupt, even though they have more assets than liabilities.

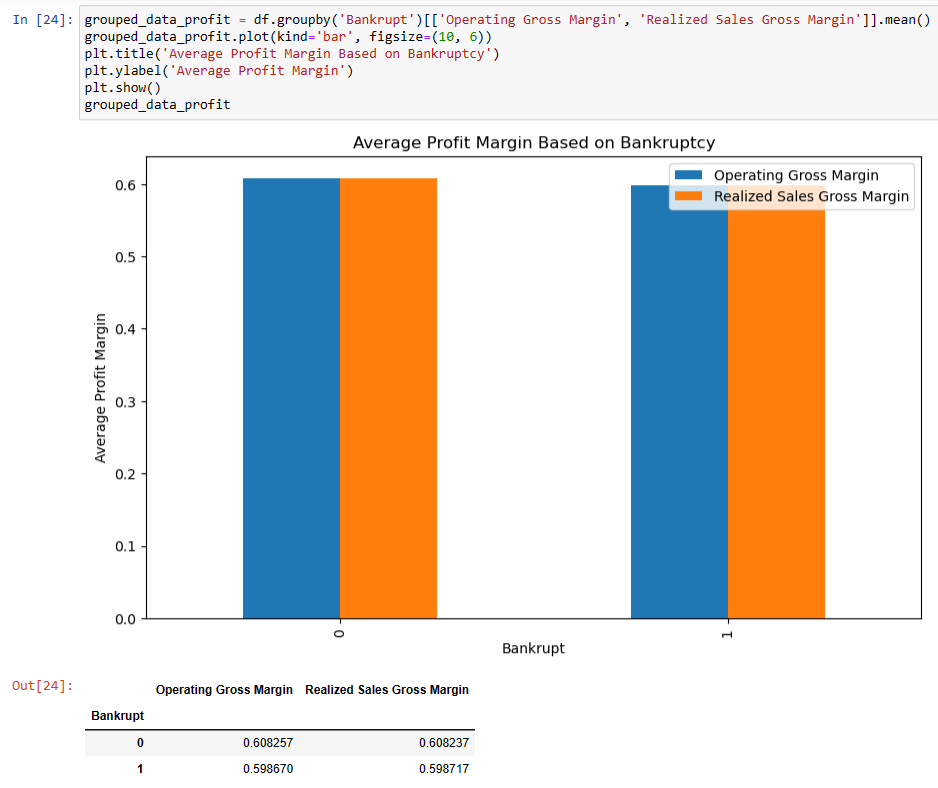


'Net Income Flag' shows the income status of the organization in the last two years, where if the net income for the last two years is negative, the flagged value is 1, otherwise the value is 0. In the dataset it shows everything is loss for the last two years.

Many organizations that experienced losses over the past two years have stabilized their businesses and thus avoided bankruptcy.

**Average Profit Margin Based on Bankruptcy:**

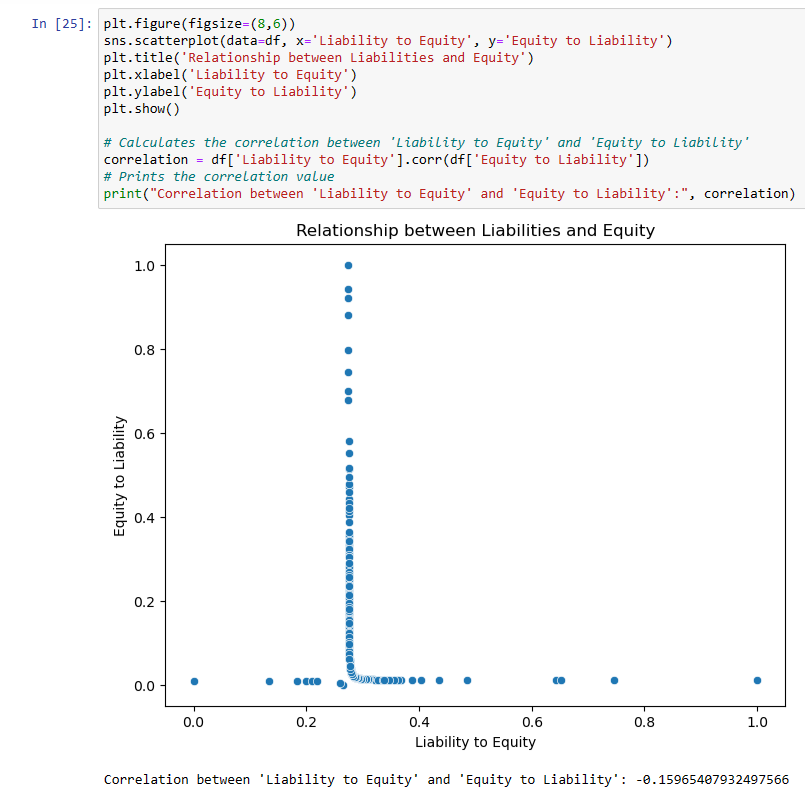
We had grouped the data based on the 'Bankrupt' label and calculate the average for the profit margin feature. This can provide insight into how profit margins relate to the likelihood of bankruptcy.

****

From a business perspective, the data shows that the average 'Operating Gross Margin' and 'Realized Sales Gross Margin' between the two classes, namely the non-bankrupt class and the bankrupt class, has very small differences. This means that on average, the operating profit margin and sales profit margin generated by non-bankrupt companies and bankrupt companies are almost the same.

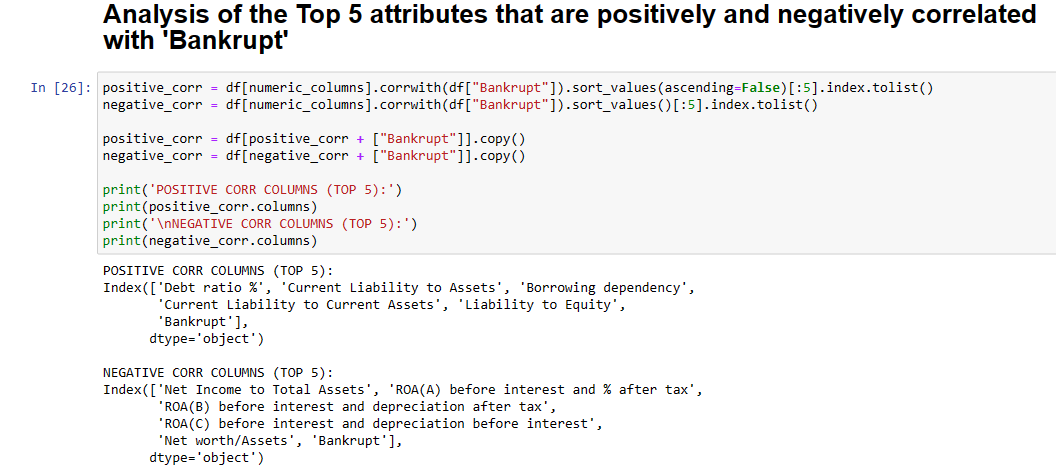
**Correlation between the liabilities and equity:**

We had used scatter plot to know the relation between these.

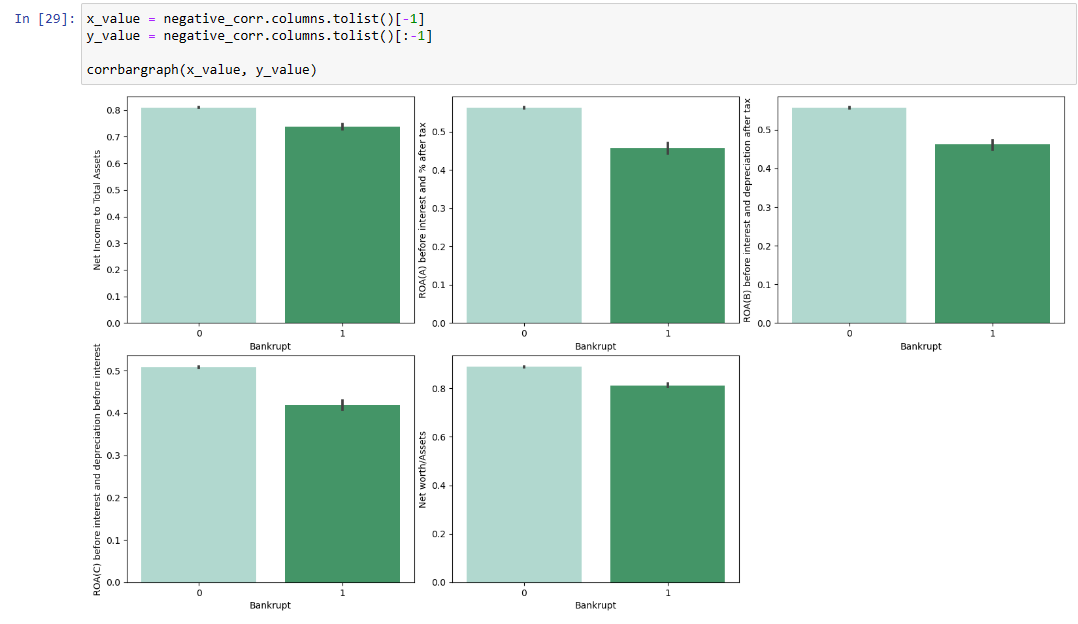


There is a negative correlation between both liabilities and equity which is -0.15965407932497566.

In a business context, this could mean that there is little relationship between a company's liabilities and its equity, and changes in one of the two do not significantly affect the other. In other words, the company's obligations (liabilities) do not seem to be very related to the company's equity.



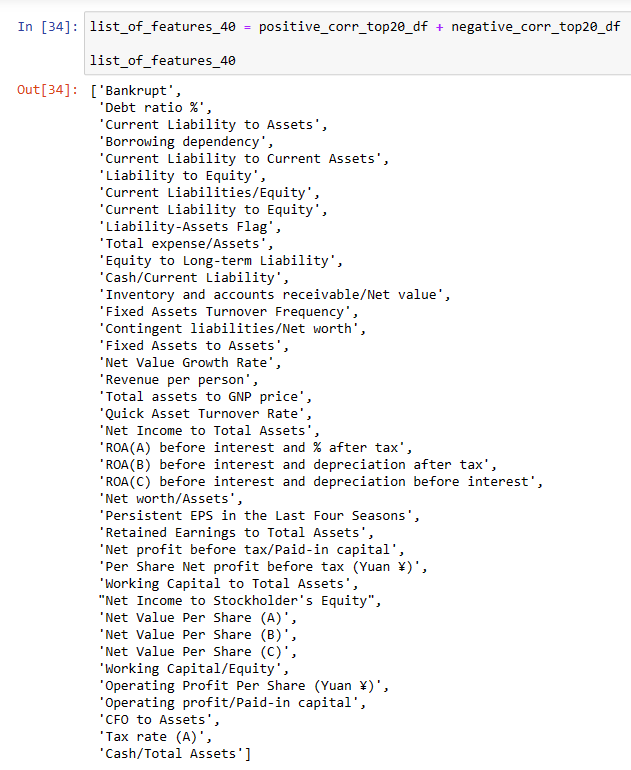




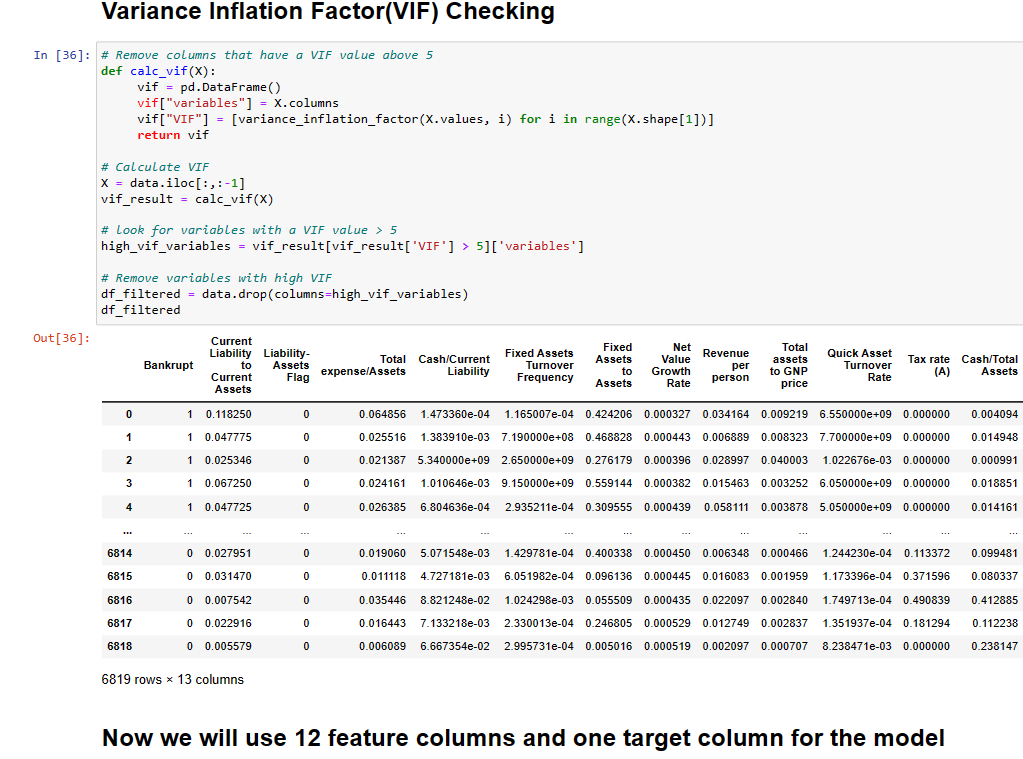
We see that three attributes – “Debt Ratio %, Current Liability To Assets, Current Liability To Current Assets” are generally high in bankrupt organizations. These attributes show us that the more assets and revenues a company has, the less likely the organization is to go bankrupt.

**Findings in Data Analysis:**

* The number of organizations that have gone bankrupt in the 10 years between 1999 - 2009 is small.
* Some companies have a lot of assets, which is always a good sign for an organization.
* An organization cannot guarantee not to go bankrupt, even though it has many assets.
* The organization in the dataset experienced losses for the last two years because its net profit was negative.
* Very few organizations that had negative revenues in the last two years went bankrupt.
* It is observed that the attributes "Debt Ratio %, Current Liability To Assets, Current Liability To Current Assets" are some of the attributes that have a high positive correlation with the target attribute 'Bankrupt'.
* An increase in the value of the attribute "Debt Ratio %, Current Liability To Assets, Current Liability To Current Assets" causes the organization to experience large losses, which ultimately results in bankruptcy.
* Increasing the value of attributes that have a negative correlation with the target attribute helps organizations avoid bankruptcy.
* It can be seen that there is a relationship between attributes that have a high correlation with the target attribute and a low correlation with the target attribute.
* We observed some correlations among the top 10 attributes, one of which is “Net Worth/Assets and Debt Ratio %” which are negatively correlated with each other.

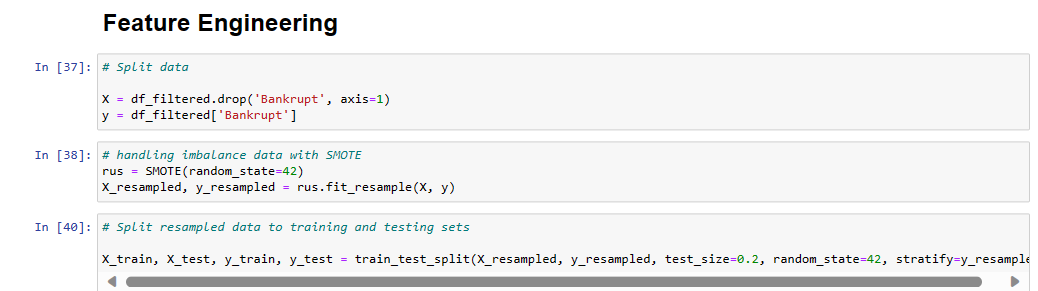


Here is the list of 40 variables which consists of top 20 positive correlated and top 20 negative correlated variables.



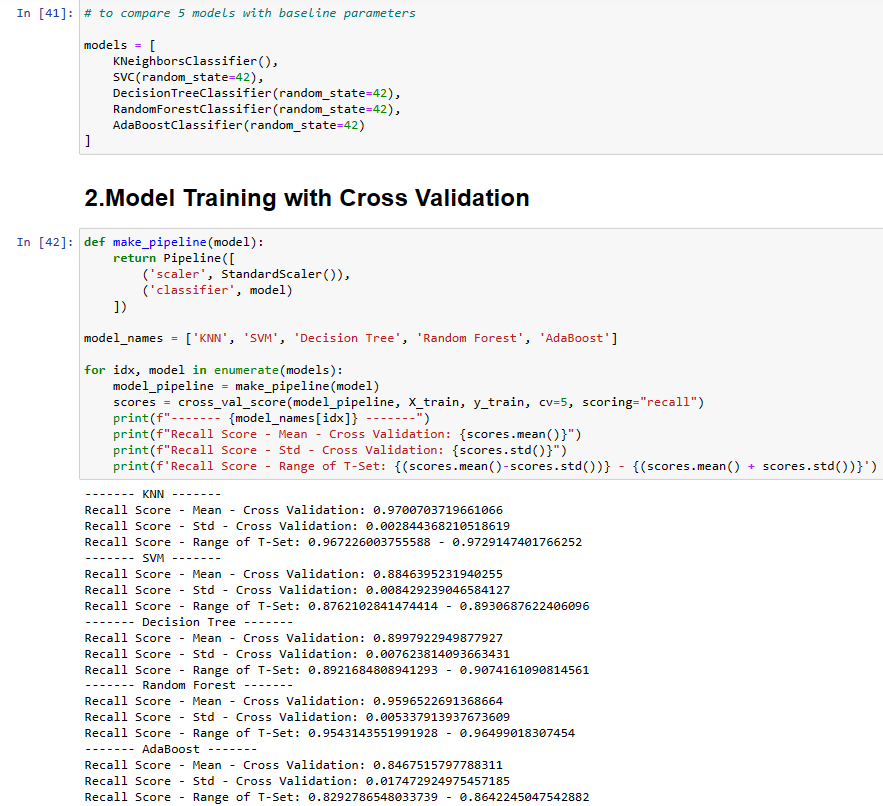
We had performed a Variance Inflation Factor(VIF) process on the top 40 variables to avoid the multicollinearity between the independent variables. We took a condition that if the VIF > 5 then the columns will be deducted from the dataframe. After the complete VIF check, We will be using 12 feature columns(x variables) and one target column( Y variable) for the model.

**Splitting the data into Training and Testing sets:**

****

Here, we had split the data into training and testing data as x\_train,x\_test,y\_train,y\_test.

**Defining the Model**

****

Here we are comparing the sets with 5 models(KNN, Decision Tree, Random Forest, Adaboost Classifier, SVC) and will select the best suitable model for the data.

**Justification for choosing the Random Forest model**

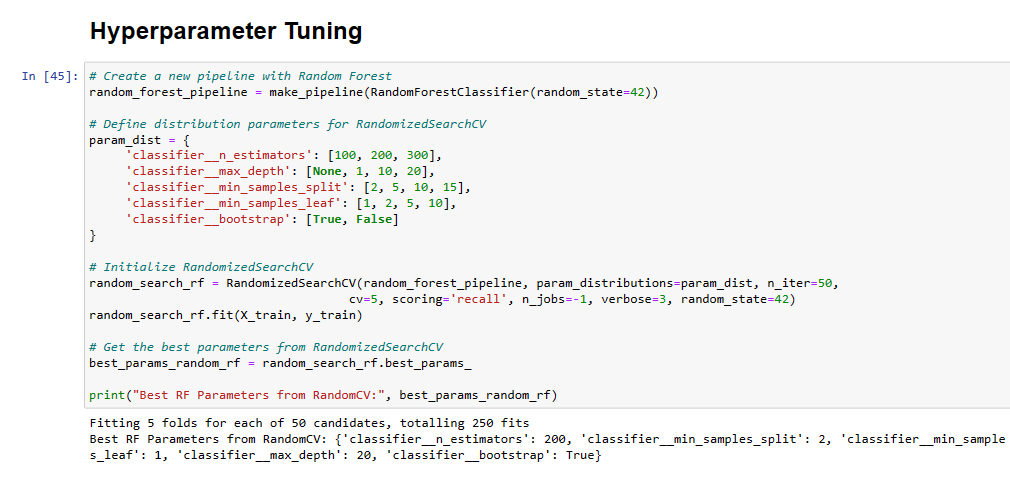
1. **Accuracy & Stability**: Random Forest shows a high recall score (0.9597) with a low standard deviation (0.0053), indicating good and consistent performance.
2. **Overfitting Prevention**: Compared to single decision trees, Random Forest is more resistant to overfitting thanks to its ensemble approach.
3. **Feature Interpretability**: Despite being an ensemble model, Random Forest can provide information about feature importance, making it easier to understand which features contribute most significantly.
4. **Flexibility**: Able to handle categorical and numeric features as well as data imbalances well.
5. **Scalability**: Although it requires training time, Random Forest provides fast and good predictions for large datasets.

Based on the above considerations, Random Forest was deemed suitable as a classification model for this project.

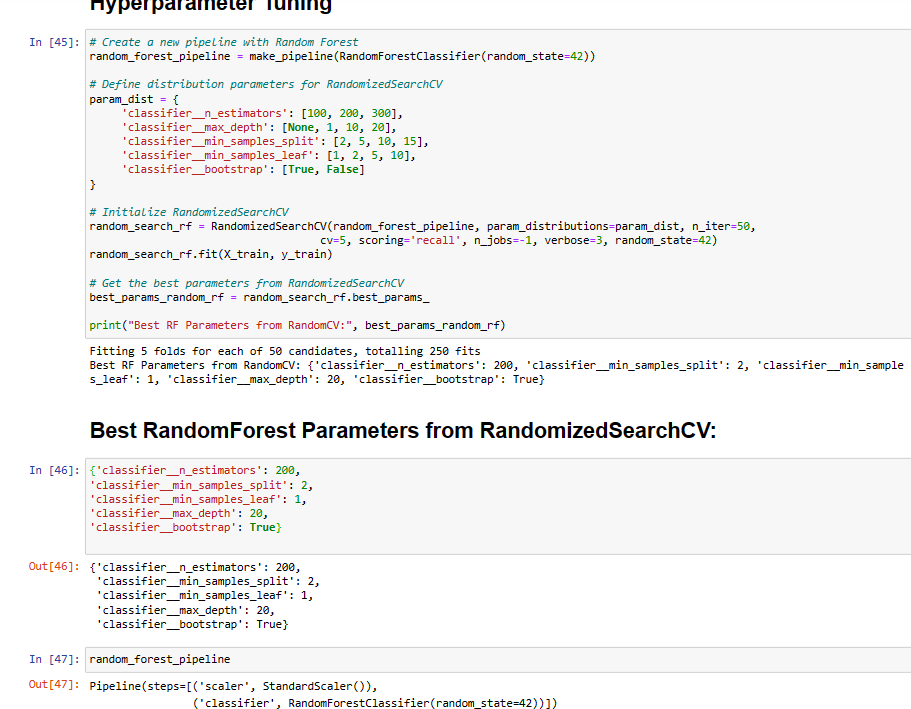
**Model Training:**

In the context of corporate bankruptcy prediction, the most relevant and important evaluation metric is recall (or sensitivity).

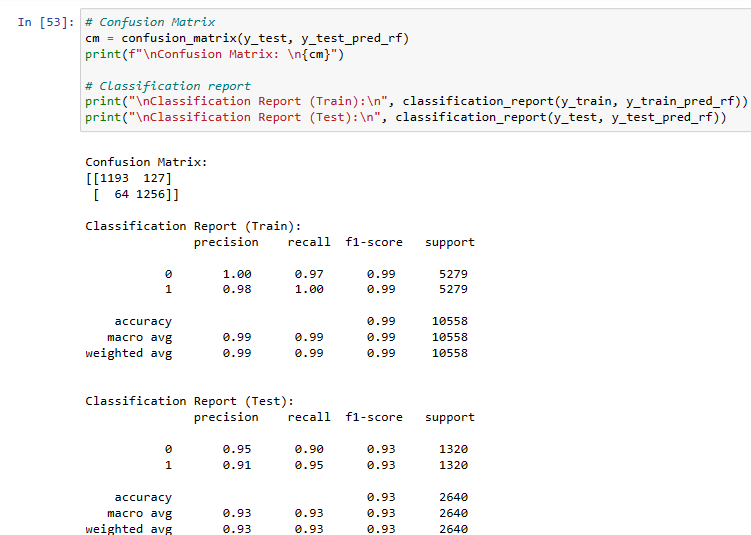
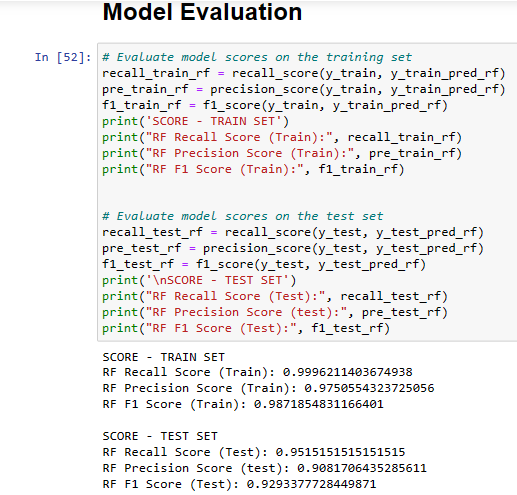
Recall measures the extent to which your model can identify companies that are truly about to go bankrupt (positive class, target 1). Due to the significant financial and business impacts associated with companies going bankrupt, we must minimize the number of “False negative” cases (companies that are actually going bankrupt but are not detected by the model) so that we can take appropriate action.

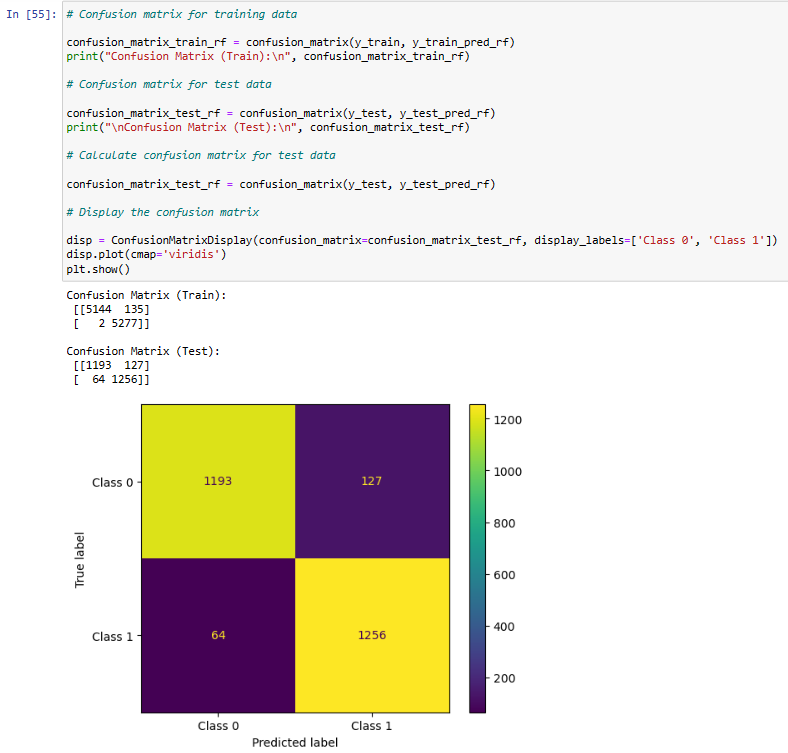
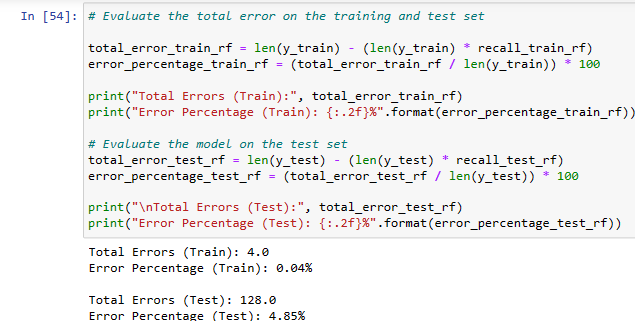


We had performed the hyperparameter tuning for the model to achieve the best accuracy.



**Model Evaluation:**

****

****

**Random Forest Model Evaluation Results:**

**Recall Score**:

* On the training data, this model achieved a recall score of 0.9996, which is almost perfect. This shows that almost all positive classes can be identified correctly by the model.
* In the test data, the recall score is 0.9515, which means 95.15% of the positive classes were correctly identified by the model.

**Confusion Matrix**:

* On training data: Of the 5279 actual observations with label 0, 5144 of them were correctly classified by the model (True Negative) while 135 were incorrectly classified as class 1 (False Positive). Of the 5279 actual observations with label 1, 5277 of them were correctly classified by the model (True Positive) while 2 were incorrectly classified as class 0 (False Negative).
* On test data: Of the 1320 actual observations with label 0, 1193 of them were correctly classified by the model (True Negative) while 127 were incorrectly classified as class 1 (False Positive). Of the 1320 actual observations with label 1, 1256 were correctly classified by the model (True Positive) while 64 were incorrectly classified as class 0 (False Negative).

**Classification Report (Test Data)**:

* Class 0: Precision 0.95 and Recall 0.90. Of all the class 0 predictions made by the model, 95% were actually class 0. Meanwhile, of all the actual class 0 observations, the model succeeded in identifying 90% of them.
* Class 1: Precision 0.91 and Recall 0.95. Of all class 1 predictions by the model, 91% were actually class 1. Meanwhile, of all actual class 1 observations, the model succeeded in identifying 95% of them.
* The F1 score for both classes is 0.93, indicating a good balance between precision and recall.
* The overall accuracy of the model on the test data is 0.93, which means 93% of the model predictions are correct, while 7% are incorrect.

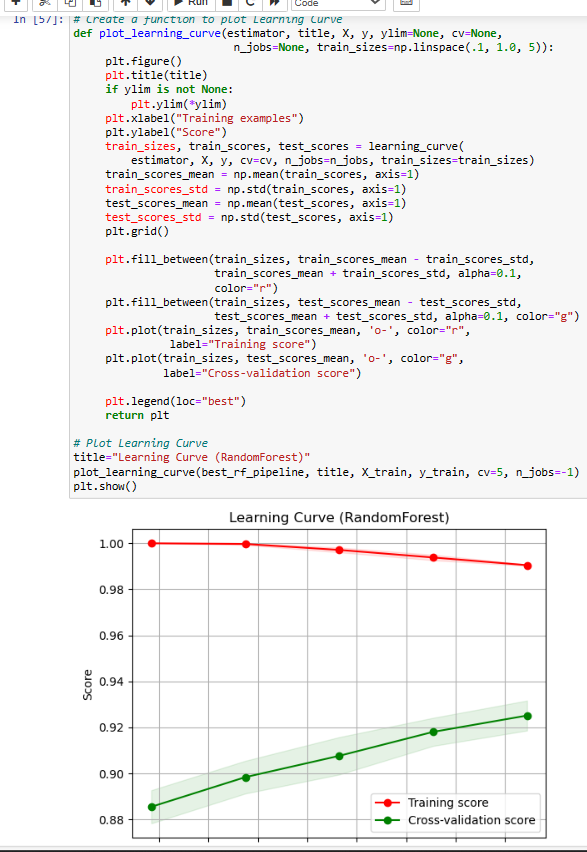
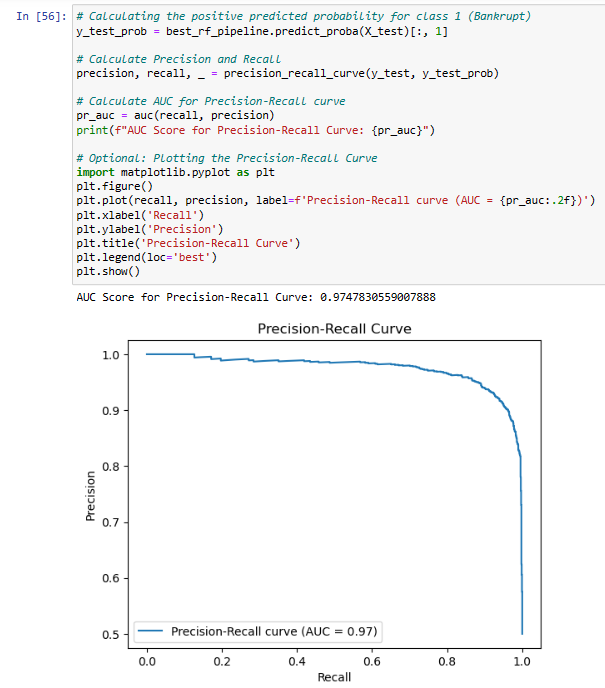
**Error**:

* The total error made by the model on the training data is 4 out of 5279 observations, which is 0.04% of the total training data.
* The total error made by the model on the test data is 128 out of 2640 observations, which is 4.85% of the total test data.

**Conclusion of Random Forest Results Model:**

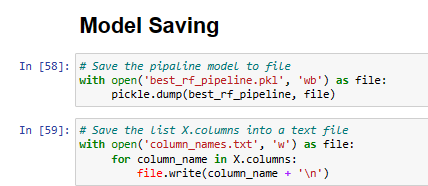
The Random Forest model I developed showed excellent performance with an accuracy of 93% on test data. Although a very high recall score on the training data may indicate potential overfitting, the consistently high performance on the testing data indicates that the model has a good capacity to generalize and is effective in classifying unknown data.

**Area Under the Curve(AUC)**

****

The AUC (Area Under the Curve) score of the model is 0.978. The AUC score ranges from 0 to 1, where a score of 1 indicates that the model has perfect classification ability, while a score of 0.5 indicates that the model has classification ability equivalent to random chance. With an AUC score close to 1, namely 0.98, this shows that the model has excellent classification capabilities and can differentiate well between the two existing classes with a high level of confidence. This indicates good model performance in differentiating between classes.

**Model Saving:**

****

Saving the model to make it ready for the deployment in the real life situations.

**Conclusion:**

**General description**:

The main purpose of this analysis is to help the an Investment Group, especially in the stock market, to make the right decision to invest in a company by making decisions based on the risk of bankruptcy.

**Model Analysis Results**:

1. The Random Forest model I developed showed excellent performance with a recall score of 0.9515 on test data. This demonstrates the model's ability to identify companies at risk of bankruptcy with a high level of confidence.
2. With an AUC score of 0.978, this model can differentiate between safe and risky companies very well. An AUC score close to 1 confirms the reliability of the model in differentiating bankruptcy risk at various thresholds.
3. Based on the important features of the Random Forest model, the model can identify which financial or operational factors contribute most to bankruptcy risk. This will help companies prioritize which areas need attention to reduce their risks.