

**Logistic Regression Project on**

**Company Bankruptcy Prediction**

**Project performed by**

**Sunil Kumar**



**Table of Contents**

* Summary............................................................................1)
* Problem statement..............................................................2)
* Introduction.........................................................................3)
* Technical Work...................................................................4)

1. Library used.
2. Function and Method used.

* Step involved in exploratory data analysis..........................5)
* Procedure performed and Observations.............................6)

**Summary:**

A company faces bankruptcy when they are unable to pay off their debts. The Taiwan Economic Journal for the years 1999 to 2009 has listed the details of company bankruptcy based on the business regulations of the Taiwan Stock Exchange. The Taiwan Stock Exchange was established in 1961 and began operating as a stock exchange on 9 February 1962. It is a financial institution located in Taipei, Taiwan. It has over 900 listed companies. The data includes a majority of numerical attributes that help understand the possibility of bankruptcy.

**Problem Statement**

The goal of this project is that you will use various classification algorithms on bankruptcy dataset to predict bankruptcies with satisfying accuracies long before the actual event.

**Technical Work**

* ***Library used***
* ***Functions and methods used.***

**Library used**

* **Pandas** This library is used in data analysis and manipulation and importing files.
* **Matplotlib** Used infor data visualization, graphs & plotting.
* **Seaborn** Used inmaking data visualization in a more colorful and meaningful way.
* **Sklearn:** useful and robust library for machine learning in Python.
* **Imblearn:** Imblearn techniques are the methods by which we can generate a data set that has an equal ratio of classes.
* **Voting classifier:**  A Voting Classifier is a machine learning model that trains on an ensemble of numerous models and predicts an output (class) based on their highest probability of chosen class as the output.
* **os** Used in provides functions for creating and removing a directory folder.

**Common functions and methods used**

**pd.read\_csv():**used in to import data in CSV format. This function has a number of arguments, but the only essential argument is file, which specifies the location and filename.

**%cd :** magical command finds present working directory.

**os.chdir():** Changes your directory to your data directory.

**head():** Used in checking the first rows of the dataset.

**tail():** Used in fetching the last rows of the dataset.

**shape():** Used in getting the total no. Of columns and rows, a dataset contains.

**Type():** Used in knowing the data structure of the dataset.

**df.dtypes:** Used in understanding the data types of the dataset.

**IsNull():** used in checking if any null values are available.

**Sum():** used in giving sum of the values.

**value\_counts() :** Used in providing counts of particular values present in data.

**describe()** : This method is used for calculating some statistical data like percentile, mean and std of the numerical values of the Series or DataFrame. It analyzes both numeric and object series and also the DataFrame column sets of mixed data types.

.**corr():** aggregate function returns a coefficient of correlation between two numbers.

**unique():** function is used to find the unique elements of an array.

**Steps involved:**

* *Setting directory/path.*
* *Loading dataset.*
* *Data cleaning.*
* *Data wrangling.*
* *Visualizations.*
* *Data pre-processing- feature selection..*
* *Modelling.*
* *Performance improvement.*
* *Model selection and evaluation.*
* *Deployment.*

**Setting directory/path**

Before we proceed with python to understand the data, it is important to let python understand first that where our data available is so that, we can perform the operation on data using python. The setting directory is similar to that and helps in understanding the presence of the data.

**os.chdir(your path)**

**Loading dataset**

Once the directory is set now we need to load the dataset which is commonly done using:

**Pd.read\_csv(‘file\_name’)**

**Data cleaning**

Now once we have our dataset loaded it is important that we look for data in good format and it is not containing anything which is not required for our analysis or which contains values that make no sense.

Data that has null values in this step of modeling, treating the null values or empty columns in this process is called data cleaning.

Essentially data cleaning has below operations to perform:

1. Dropping and selecting columns in a DataFrame.
2. Changing the Index of a DataFrame.
3. Tidying up Fields in the Data.
4. Cleaning the Entire Dataset in terms of null values.
5. Renaming columns and skipping rows.
6. Looking for duplicates.
7. Searching for outliers and treating accordingly.

**Data wrangling**

Process of cleaning and unifying messy and complex data sets for easy access and analysis. Converting the format of data to a suitable format.

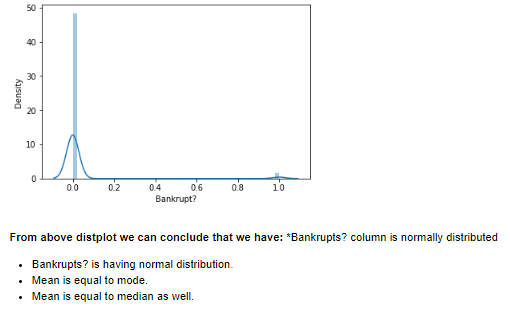
We can rename our variables using the .rename() function put your columns inside a list of a dictionary in case columns are more and which requires renaming to a more sensible form.

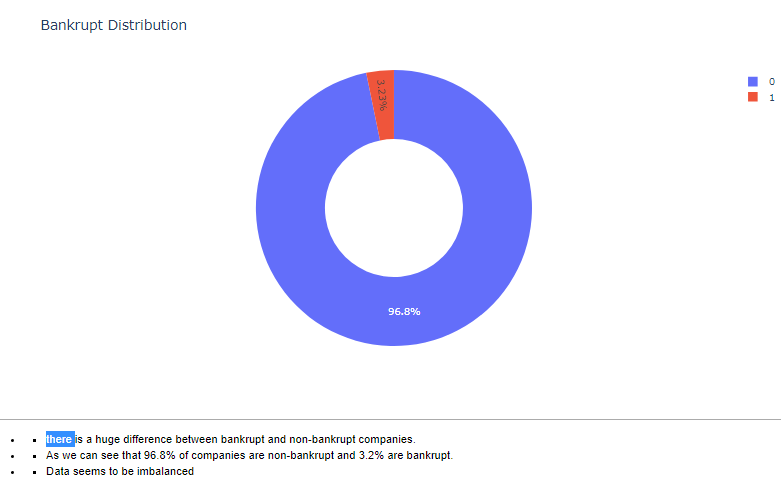
This step involves finding manually the data and making a suitable alteration for that it make more sense.

**Data Visualization**

This step of visualization involves checking the data in pictorial form. We can see outliers explore null values and also we can understand the variables which are correlated to each other’s and how data is distributed.

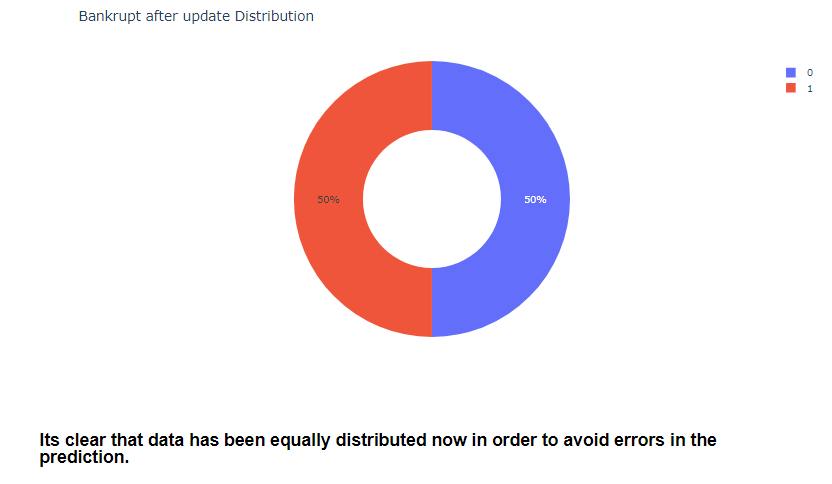
Apart from this, visualization helps in easy understanding of the data and representation of most of the data information in right and quick fashion.





This data shows that we have 96.8% of observations which are not bankrupt and we have only 3.23% of observations which are bankrupt.

Since this seems to be a case data imbalanced and would required balancing before we try modeling on this data.



After using suitable re-sampling techniques data now seems to be having balanced classes.

**Modelling**

Before we proceed with model building, it is important that we have the right amount of data and we understand the features/variables very well. Almost every model takes the numerical data for further processing in the model part so it is important to convert the entire data into numerical columns so that we can process it further.

It is equally important to classify your columns into dependent and independent variables. Once this task is performed we are ready for the next step in modelling which can be splitting the training and testing data.

 Again if there is the case that we have only imbalanced classes like data it should be treated accordingly like using the right techniques of re-sampling while ensuring not to lose on data information, so it’s equally important to use the right re-sampling technique when that feature we have is greatly impactful or have greater significance for our results.

 Next in this process, we would need some of the machine learning libraries and suitable models for the prediction since in this case we are predicting the bankruptcy of a company which is a binary classification problem and we can start with a simple model like Logistic regression, its uses the sigmoid function and predicts the probability of the actual class.

After fitting the data look for the scores if we need to improve the scores, look for the scope of improvement if required we can use regularization, Hyperparameter tuning or in this case of classification we can use ensembles techniques like boosting and bagging and Finally Model has to undergo a thorough evaluation which checks whether your model is doing good in terms of prediction in this case of classification evaluation parameter can be accuracy, precision, Recalls, F1 score, and AUC-RUC curve.

**Model Used**

Logistic Regression.

Decision tree.

Random forest.

AdaBoost.

Gradient Boost.

Voting Classifier

**Procedure performed and Observations**

Started with data loading and importing the libraries and then started with exploring the data and looking into columns and rows. It was seen that there were no missing values.

An organization cannot guarantee not being bankrupt, although owning several assets.

The organizations in the dataset are running into losses for the past two years as their net income poses a negative increase in the values of the attributes that have a negative correlation with the target attribute helps an organization avoid bankruptcy. There seems to be a relation between attributes that have a high correlation with the target attribute and a low correlation with the target attribute.

I observed several correlations among the top 12 attributes, one of which being “Net Worth/Assets and Debt Ratio %” that is negatively correlated with one another.

 While looking into Bankrupt? Column it was seen that column was following an almost normal distribution. While exploring Bankrupt? Column further it was evident that data is imbalanced and there is a huge difference between bankrupt and non-bankrupt companies.

I have used 2 techniques to overcome this problem. 1st was to create a data frame and divide it into equal rows that were equal (220,220) rows and 96 columns.

Once I was done here I then applied modelling techniques started with Logistic regression and the other ensembles.

I achieved consistently better quality with a better model every time, I also have used a voting classifier which basically takes an ensemble of numerous models and gives the best-predicted output/accuracy. The voting classifier has given me the best quality. 2nd I have used Synthetic Minority Oversampling Technique (SMOT) in which I oversampled the minority class and balanced the data. I also performed the normalization on non-fractional columns to make sure data followed the same scale. I have seen and found that after SMOT results for accuracy were better than the 1st technique.

Finally, I have used calibration, it basically assures the reliable benchmark and accuracy results when features are very important, by this way I have achieved the best accuracy for the model prediction.

I have seen that the Random Forest classifier and voting classifier worked exceptionally well.

**Challenges.**

The major challenges I have faced in this project are mentioned below:

I handled the imbalanced data.

The non-fractional column contains uneven values.

Choosing the right model for this problem, explored the Voting classifier.

While oversampling I was unable to create a data frame and results were coming out different even after a few iterations.

Calibration of the Model.