Assignment 2

MCA Technology Solutions Private Limited was established in 2015 in Bangalore with an objective to integrate analytics and technology with business. MCA Technology Solutions helped its clients in areas such as customer intelligence, forecasting, optimization, risk assessment, web analytics, and text mining, and cloud solutions. Risk assessment vertical at MCA technology solutions focuses on problems such as fraud detection and credit scoring. Sachin Kumar, Director at MCA Technology Solutions, Bangalore was approached by one of his clients, a commercial bank, to assist them in detecting earning manipulators among the bank's customers. The bank provided business loans to small and medium enterprises and the value of loan ranged from INR 10 million to 500 million ($1 = INR 66.82, August 16, 2016). The bank suspected that their customers may be involved in earnings' manipulations to increase their chance of securing a loan.

Saurabh Rishi, the Chief Data Scientist at MCA Technologies, was assigned the task of developing a use case for predicting earning manipulations. He was aware of models such as *Beneish* model that was used for predicting earning manipulations; however, he was not sure of its performance, especially in the Indian context. Saurabh decided to develop his own model for predicting earning manipulations using data downloaded from the Prowess database maintained by the Centre of Monitoring Indian Economy (CMIE). Daniel received information related to earning manipulators from Securities Exchange Board of India (SEBI) and the Lexis Nexis database. Data on 220 companies was collected to develop the model.

The data is available in the file *"Earnings Manipulation 220.csv"* in the following link.

https://drive.google.com/file/d/1h-qteu-obJ5y5nQhHn_ypUOzZSp8lEVn/view?usp=sharing

Description of the columns is given in the following Table. In this table

- netPPE is net Property, Plant, and Equipment. Property Plant and Equipment is the value of all buildings, land, furniture, and other physical capital that a business has purchased to run a business.
- Subscript (*t*) means value in the current year financial statement and subscript (*t* – 1) means the value in the previous year financial statement.

Column	Description
Company Name	It is a serial number.
Year Ending	The year to which the financial statement belongs to.
Days Sales to Receivables Index (DSRI)	$\mathrm{DSRI} = \dfrac{\dfrac{\text{Receivable}_{(t)}}{\text{Sales}_{(t)}}}{\dfrac{\text{Receivable}_{(t\text{-}1)}}{\text{Sales}_{(t\text{-}1)}}}$
Gross Margin Index (GMI)	$\mathrm{GMI} = \dfrac{\dfrac{\text{Sales}_{(t-1)} - \text{Cost of Goods Sold}_{(t-1)}}{\text{Sales}_{(t-1)}}}{\dfrac{\text{Sales}_{(t)} - \text{Cost of Goods Sold}_{(t)}}{\text{Sales}_{(t)}}}$
Asset Quality Index (AQI)	$\mathrm{AQI} = \dfrac{\dfrac{1-(\text{Current Asset}_{(t)} + \text{netPPE}_{(t)})}{\text{Total Assests}_{(t)}}}{\dfrac{1-(\text{Current Asset}_{(t-1)} + \text{netPPE}_{(t-1)})}{\text{Total Assests}_{(t-1)}}}$
Sales Growth Index (SGI)	$\mathrm{SGI} = \dfrac{\text{Sales}_{(t)}}{\text{Sales}_{(t-1)}}$
Depreciation Index (DEPI)	$\mathrm{DEPI} = \dfrac{\dfrac{\text{Depreciation Expense}_{(t-1)}}{(\text{Depreciation Expense}_{(t-1)} + \text{netPPE}_{(t-1)})}}{\dfrac{\text{Depreciation Expense}_{(t)}}{(\text{Depreciation Expense}_{(t)} + \text{netPPE}_{(t)})}}$
Sales and General Administrative (SGAI)	$\mathrm{SGAI} = \dfrac{\dfrac{\text{SGAIExpense}_{(t)}}{\text{Sales}_{(t)}}}{\dfrac{\text{SGAIExpense}_{(t-1)}}{\text{Sales}_{(t-1)}}}$
Accruals to Asset Ratio (ACCR)	$\mathrm{ACCR} = \dfrac{\text{Profit after Tax}_{(t)} - \text{Cash from Operations}_{(t)}}{\text{Total Assests}_{(t)}}$
Leverage Index (LEVI)	$\mathrm{LEVI} = \dfrac{\dfrac{(\text{LTD}_{(t)} + \text{Current Liabilities}_{(t)})}{\text{Total Assests}_{(t)}}}{\dfrac{(\text{LTD}_{(t-1)} + \text{Current Liabilities}_{(t-1)})}{\text{Total Assests}_{(t-1)}}}$
MANIPULATOR	1 – Company has manipulated the financial statement (Manipulator) 0 – Company has not manipulated the financial statement (Non-Manipulator)

Answer the following questions using the above dataset.

1. How many cases of manipulators versus non-manipulators are there in the dataset? Draw a bar plot to depict.

2. Create a 80:20 partition, and find how many positives are present in the test data.

3. The number of cases of manipulators are very less compared to non-manipulators. Use upsampling technique to create a balance dataset.

4. Build the following models using balanced dataset. Comment on which metric should be given preference for this dataset. Finalize the model for each technique after Hyperparameter tuning using GridsearchCV based on the metric selected. Compare the model performances with respect to different evaluation metrices.

    a. Naïve Bayes
    b. KNN
    c. SVM
    d. Logistic regression
    e. Random Forest
    f. Adaboost
    g. Gradientboost
    h. XGBoost

5. Comment on which are the most important features for predicting the manipulators.

6. The number of cases of manipulators are very less compared to non-manipulators. Use downsampling technique to create a balance dataset.

7. Compare the results of using both upsampling and downsampling techniques. Report the best model of all the models.

What will you have to submit?

You need to submit a single jupyter notebook (.ipynb) file with all the work done including comments via the elearn(Taxila) portal. Naming convention of the file should be <BITS-ID-number>.ipynb.

Deadline:

19th November, 2023

Note that this is a hard deadline and no extension will be granted further. Copying efforts will be dealt very seriously, students involved, if caught, will be given zero marks.