Date __/__/__
Page _____
STUDY BUDDIES

Name → Mudit Vyas : idmvyas@official@gmail.com
Topic → CORPORATE BANKRUPTCY PREDICTION

★ Objective : To identify the best classification model in terms of accuracy & performance for predicting the bankruptcy of corporations using various statistical forecasting techniques : Logistic Regression, SVM, NN & BNB.

★ Introduction : ① Bankruptcy is a legal proceeding involving a person / business that is unable to repay their outstanding debts.

② Bankruptcy Prediction is the art of predicting bankruptcy & various measures of financial crisis or distress of public firms.

③ Significance : Predicting financial distress is to develop a reliable model that will provide measure & predict the financial condition of a corporate entity.

Causes for Corporate Bankruptcy (possibilities) :

① Market Conditions :⇒ Poor conditions in overall economy & the specific market in which business operates are common causes of bankruptcy.
⇒ Competition from larger companies is another market factor that can cut into revenue of small companies & lead to bankruptcy.

② Financing : Many business owners take out loans to help finance their operations. If business struggles, his lender may not willing to grant additional funding leading to bankruptcy.

③ Poor Decision Making

④ Others : Location, loss of key employees, competitions & legalities

— x — x —

(2)

Date __/__/__
Page _____
STUDY BUDDIES

★ Datasets & features

⇒ Data is available in 5 cases depending ~~upon~~ forecasting period.

(I) 1st year forecasting period: Contains 7,027 financial statements. (5yrs)

7,027 Statements
- ▷ 271 or 3.8% : Bankrupted Companies
- ▷ 6,756 or 96.2% : Not Bankrupted companies

(II) 2nd year Indicates bankruptcy status ~~after~~ after 4 yrs

10,173 Statements
- ▷ 400 (3.9%) Bankrupted
- ▷ 9,773 (~96%) Not Bankrupted

Companies

(III) 3rd Year : Similar to 2nd year interval of forecasting

10,503 Statements
- ▷ 495 (4.7%) Bankrupted
- ▷ 10,008 (~95%) Not Bankrupted

(IV) 4th Year

9,792 Statements
- ▷ 515 (5%) Bankrupted
- ▷ 9277 (95%) Not Bankrupted

(V) 5th Year : Indicates bankruptcy status after 1 year

5,910 Statements
- ▷ 410 (6%) Bankrupted
- ▷ 5500 (94%) Not Bankrupted

Dataset : UCI Machine Learning Repository

(5)

★ Features : In our dataset, we have 64 features indicating financial health of the corporate entity and one label column to indicate bankruptcy status after 3 years/5 yrs

★★ Missing Data : Impact Accuracy & Efficiency of our model. Dropping such data is likewise harmful. We can use statistical techniques like Mean, Median, Nearest Neighbours & Multivariate Imputation.

★ Machine Learning Models & Training Algorithms :

» Known Labels : Bankruptcy flag → 0 or 1
0 means Not Bankrupted Company
1 means Bankrupted Company.
⟹ Supervised Machine Learning + Classifier ~~technique~~ model

(i) Logistic Regression : Linear model of Regression + Use of L2 regularization to avoid overfitting of data & introducing high Variance in our model.
Sigmoid Function → Probabilities describing the possible outcomes are modelled using a logistic fn^c.

$$p(y|x;\theta) = (h_\theta(x))^y (1 - h_\theta(x))^{1-y}$$

Using Scikit-learn, cost fun^c mention is regularized by :

$$\min_{\omega, c} \frac{1}{2}\omega^T\omega + C\sum_{i=1}^{n} \log(\exp(-y_i(X_i^T\omega + c)) + 1)$$

(ii) Support Vector Machine : Constructs a hyper plane or set of hyper planes in a high dimensional space, which can be used for classification task.

(9)

Advantage of SVM : ① Effective in high dimensional space.
② Also uses a subset of training pts in decision func (called support vectors)
Our cost func has below:

$$-\sum_{i=1}^{n} a_i y^{(i)} \langle x^{(i)}, x \rangle + b$$

③ Neural Network : (a) Multi layer Perceptron Algorithm that trains using backpropagation.
(b) Adam Gradient descent → Weights & Constant optimize
(c) Square Error Loss func
(d) 1st dense layer → n × hidden layer → (2 neuron) Output layer

④ Naives Bayes : Based on bayes theorem with 'naive' assumption of conditional independence b/w every pair of features given the value of the class variable

Bayes theorem States : given class variable & dependent feature vector

$$P(y|x_1, \ldots, x_n) = \frac{P(y)\,P(x_1, \ldots, x_n|y)}{P(x_1, \ldots, x_n)}$$

⑤ Extreme Gradient Boosting :
Based on the principle of gradient boosting Framework.
Gradient boosting ~~produces~~ produces a predictive model a model in the form of an ensemble of weak prediction models

We use a depth of 3 with squarred error objective func & gblinear booster with L2 regularization.

(6) Light Gradient Boosting Machine : It is gradient boosting framework that uses tree based learning algorithm & optimizes using histogram based algorithm for performance efficiency.

Results : Learning Rate of L_2 regularized method : 0.001
We found that any other learning rate was causing delay in converging.

⇒ Cross Validation : Performing K-Fold cross validation for evaluation the performance of our models where K = 5.

⇒ Metrics : Primary metrics used to evaluate the performance of models are Accuracy, Score, log Loss, Fit Time Score & confusion matrix over both train & test dataset.

Observation : (i) SVM performed better than any other model in terms of maximum Accuracy & minimum loss.

(ii) Accuracy Score for Test Score :
SVM (94.7 %) ≥ NN (94.7%) > LR (94.6%) >> NB (75.58%) > LGBM (70.8%) > XGB (52.64%)

CHALLENGES :

(i) Unexpected Performance of Gradient Boosting which require inspection to ~~evate~~ elevate performance.

(ii) NB only achieved 76.54% train Accuracy which leads to less test Accuracy. (Underfitting issues)

(iii) Optimizing & Hyperparamter tuning require for NN, LR & NB.

6

ROADMAP

- Defining Problem
- Data Collection
- Data Inspection
  - Missing Data
  - Outliers
  - Noise Present
  - → Data Wrangling
- EDA
  - finding correlation of attributes to bankruptcy flag (0 or 1)
  - Analysing trend & finding important features & extract them
- Dataset split → test ratio (0.2 to 0.3)
- Model → Classifier Model Algorithm
- Evaluation + cross validation → Confusion matrix + Accuracy + test loss score
- Optimisation & Repeat → Inspection of data + Hyperparameter
- Desired test prediction → Deployment