CS 559: Machine Learning: Fundamentals and Applications
Spring 2024 Semester Project: Due on May 3, 2024

Our project is to predict whether a company will file for bankruptcy. It is a binary classification: no = 0 and yes =1. The train and test data sets have 5807 rows and 1012 rows, respectively. The total number of given features is 95. The project aims to train a model that classifies whether a company will file for bankruptcy (the target column is 'Bankrupt?'). The test data set does not have the true label. Each team will submit the predicted classes for 1012 companies to the submission file.

Despite the fact that there are no limits on the use of libraries and packages, please follow the rules described below.

1 Training Data Preparation

95 features are considered too many features. The number of features must be reduced to increase the efficiency of the model. **The maximum number of features a team can have is 50**.

1. PCA is not allowed to reduce the original dimensions. It can be used only after features are modified via extraction or engineering. More than 95% of information must be kept when PCA is performed. PCA can be a good strategy at the last stage of training the data, and it is optional.
2. An original feature can be kept if and only if non-linearity is confirmed. Otherwise, it must be feature engineered.
3. A feature can be extracted if and only if it holds a strong positive or negative correlation between other features (not to the target). The correlation coefficient $|\rho| \geq 0.95$ considers the strong correlation.
4. The normality of each feature must be confirmed.

2 Company Characterization

Understanding every company's situation is quite difficult at a given time. However, knowing the general situation will greatly help model training. The easiest approach is finding the common characteristics between similar companies, and similar companies can easily be grouped by using unsupervised learning clustering techniques.

1. Find k or more subgroups by KMeans or Gaussian Mixture where k is higher or equal to the number of members in each team. When the clustering technique is applied, the target must be dropped, and the technique must not be applied to the original feature space.
2. Identify unique or helpful characteristics in each subgroup. Use visualization techniques to present identified characteristics.
4. Keep the cluster IDs. It will be used later in Section 3.

3 Train Model 1: Stacking Method

The logistic regression is fitted on the **original train data**, and the result is presented in Table 1. Because the training data is imbalanced (see Figure 1), predicting $y = 1$ accuracy (hereafter, accuracy score) is very poor because weights on important features, for $y = 1$ case, were underestimated. The accuracy score defines how successfully the model identifies the company that files for bankruptcy as follows

data set	FF(FT)	TT(TF)	accuracy	accuracy score
train	5570(39)	5(193)	0.96	0.025
test	983(7)	0(22)	0.97	0

Table 1: Logistic Regression Result Table without any modification. The second and third columns present the number of companies' true status (y) and predicted status ($\hat{y}$) as FF - ($y = 0, \hat{y} = 0$), FT - ($y = 0, \hat{y} = 1$), TT - ($y = 1, \hat{y} = 1$), and TF - ($y = 1, \hat{y} = 0$). The accuracy presented in the fourth column is the overall accuracy, (TT+FF)/(TT+TF+FF+FT). The accuracy score in the fifth column is evaluated using Equation 1.

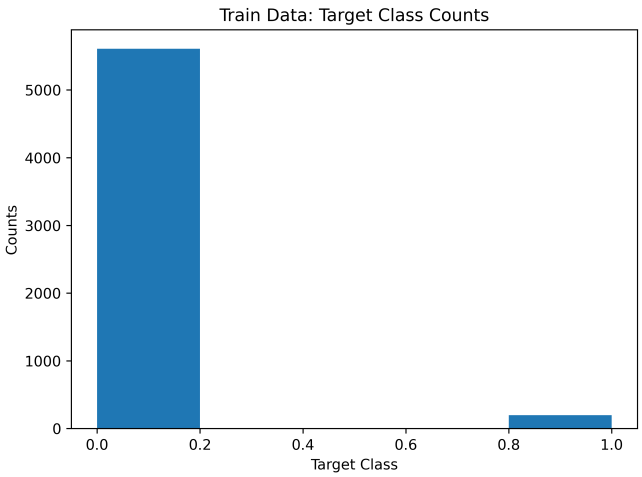

Figure 1: Train Data Target Distribution

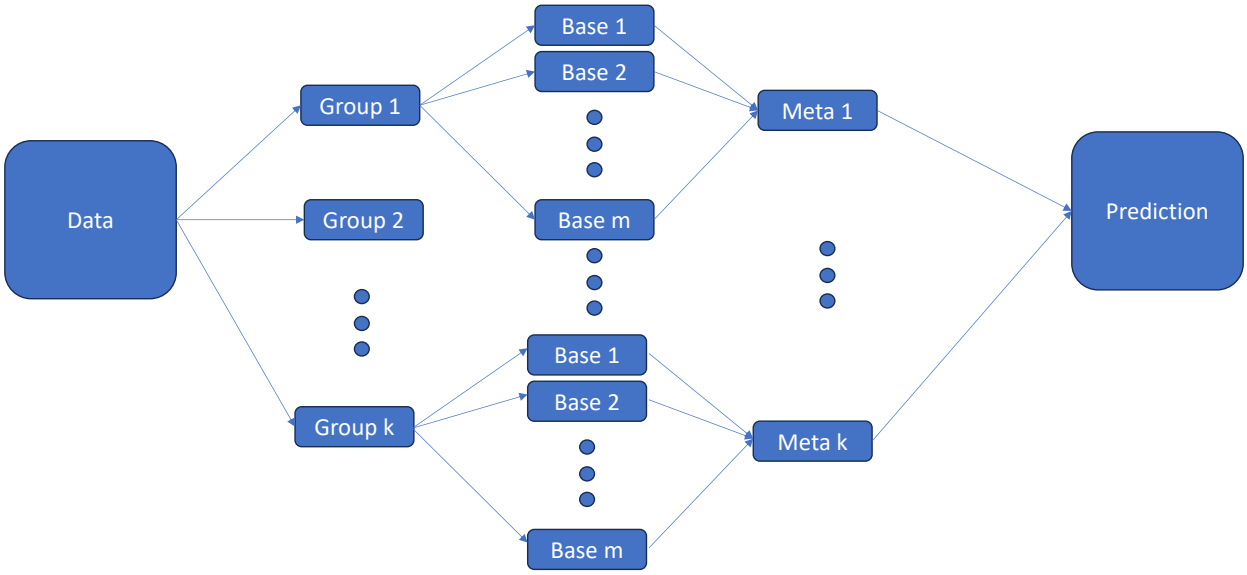

Figure 2: Stacking Method Training Model Diagram

$$acc = \frac{TT}{TF + TT} \tag{1}$$

from Table 1. Therefore, the model needs to be trained in different ways that do not underestimate crucial features among companies with $y = 1$. One of the proposed training models is the stacking method. It can be broken into two steps. The first step is to predict the group (the cluster-ID from Section 2) the company will likely belong to, and the second step is to classify whether the company will file for bankruptcy in each subgroup. Figure 2 displays the roadmap of the model.

1. Build a model that predicts a group a company will likely belong to using any supervised learning algorithm. The prediction's accuracy should be high, and it is okay to overfit. Identify the features that play important roles in this prediction.
2. Build a stacking model that predicts whether a company will file for bankruptcy.
    a. Build base models using three non-parametric models. While the features in base models must be the same, the features used in each subgroup's model do not have to be the same.
    b. Use a parametric model to build a meta-model.
    c. Each member should build a model for one or more subgroups. The member's name, the subgroup(s) worked on, and the accuracy score must be reported. Provide the confusion matrix. See Table 2 for the format.
3. Summarize the work.

4 Train Model 2: k-fold Cross Validation

An alternative way of training the model is using k-fold cross-validation. Each member will build one or more models using different supervised learning algorithms. Remember that the train data is imbalanced,

and generating each fold must be done cautiously. Choose the best model based on the accuracy score and the number of features used. Not all 50 features from Section 1 may not be used.

5 Generalization

Choose the best train model from models trained in Sections 3 and 4. Then, predict the company's bankruptcy status using the test set. Explain the reasons for choosing the best model. The predicted class and paste the result to the submission file (see Section 7.2 Table 3). The instructor will evaluate the accuracy score using the predicted class the team submitted.

6 Class Competition

This project is a class competition (Sections WS and WS1 combined), and teams will be ranked using the following metrics:

$$Rank = 0.35(acc_{train}) + 0.35(acc_{test}) + 0.3\left(\frac{50 - N_{features}}{50}\right)$$

where

- acc_{train} is the accuracy score of the train model from Section 3.
- acc_{test} is the accuracy score of generalization obtained from Section 5.
- $N_{features}$ is the number of features for the best model in Section 4.

The rank will be converted to the top percentile in the competition and will be used for the individual's project grade (see Section 7.3).

7 Timeline, Submission, Grade Scheme

The submission of the project is due on May 3rd by 11:59 PM.

7.1 Timeline

The ideal timeline of the project is broken down as follows.

1. Train Model: The train data set will be released on Monday, 3/25, so the training data preparation and building of a final model (from Section 1 to 4) can be started.
2. Generalization: **The test data will be released on the first week of Project Weeks (April 22)**. The generalization should be one-time work and should not take a long time.
3. The video presentation and files must be submitted by May 3.

7.2 Submission

Here are the requirements for submission. Each team will submit a single compressed **zip** file. The submission will require the following files inside the zip file. Name the file as *GroupNumber_CourseSection.zip*.

1. Sections 1 and 2: A single team notebook file: Clean and summarize the work. Use Markdown for comments and explanations. Name the file as *GroupNumber_CourseSection_Section2*.

Subgroup ID	Name of Student	Average accuracy score base models [TT(TF)]	accuracy score Meta model [TT(TF)]	accuracy score k-fold CV [TT(TF)]	$N_{features}$
1	John	0.93 [184(14)]	0.84 [167(31)]	0.78 [154(44)]	34
2	Kim	0.78 [$\cdots$]	0.67 [$\cdots$]	0.84 [$\cdots$]	49
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
k	Name	0.xx [$\cdots$]	0.xx [$\cdots$]	0.xx [$\cdots$]	xx

Table 2: Result Table Example: Column 1: The subgroup ID from Section 2, column 2: the name of the member who worked on the subgroup, column 3: the average of the accuracy score from beta-models in the subgroup, column 4: the accuracy score from the meta-model of the subgroup, column 5: the accuracy score from the model the member trained in Section 4, and column 6: the number of features the member used in Section 4. All the accuracy scores are as defined in Equation 1.

Index	Bankrupt?
1	0
2	1
$\vdots$	$\vdots$
1012	0

Table 3: Submission File Example

2. Sections 3 and 4: A single notebook file for individual team members. Show the work and summarize the work. **Individual files must be included in the compressed file**. If any member of a team fails to submit, the member will be excluded from the grading. Name the file as *GroupNumber_CourseSection_MemberName_Section2_3*.
3. Section 5: Submit a single team's notebook file with code and the submission file as shown in Table 3. Name the file as *GroupNumber_CourseSection_Section5*.
4. Result Table: Summarize the results in the docx file.

   – Section 2: Report the number of subgroups and their unique characteristics.
   – Section 3 and 4:: Construct a table as shown in Table 2. In the notebook file, present the length of the train data column and the confusion matrix.
   – Section 5: Explain the best model from Sections 3 and 4.
   – Section 6: Report the accuracy scores, acc_{train}, the result of the best model selected in Section 5, and the number of features $N_{features}$, used in the best k-fold cross validation model in Section 5. These values will be used for the ranking score.
   – Video Presentation: Record the video presentation to describe the workflow, models, and results. You can either include the video file in the zip file or provide the link.

   ** All reported results must be consistent with notebook files. Make sure to use **random.seed()** or **random_state=** whenever needed so the same results can be reproduced. Each notebook file must display the result. Any non-consistent results will be marked 0.

   – Name the file as *GroupNumber_CourseSection_Results*

7.3 Grade Scheme

This is how the project will be graded. Although this is a team project, each member will be graded separately, as shown below

$$Score = 0.2(Rank) + 0.4(acc_{Section3}) + 0.4\left(\frac{acc_{Section4} + (50 - N_{features,Section4})/50}{2}\right)$$

where $Rank$ is the top percentile of the competition, $acc_{Section3}$ is the accuracy score of the individual's train model in Section 3*, $acc_{Section4}$ is the accuracy score of the individual's train model in Section 4, and $N_{features,Section4}$ is the number of features the individual used in the training model in Section 4. The accuracy scores and number of features reported in Table 2 will be used. The accuracy metrics in Equation 1 will be used for the individual's grade.

Note*: If a member trained the staking model for more than one subgroup, the best-performed model in terms of accuracy will be used in the grading metric.

Note: The group performance ($Rank$) will weigh 20% of the project grade, and the remaining 80% of the grade will be evaluated from the individual's performance.

Note: This is a team project. If any member of a team is not responsive and does not return emails, texts, or calls, it is the team's responsibility to contact me. If this member was reported twice or more, the student will be out of the project and receive a zero. Please be responsive to team members.