Bankruptcy Prediction Report

Abstract

Machine learning models have increasingly been used in bankruptcy prediction. However, there is the data imbalance in historical data, which causes inaccurate prediction, resulting in huge financial losses. Therefore, a framework is used in our project to combine under sampling method and classification to address the problem.

Business Understanding

Inaccuracy in bankruptcy forecasting can negatively impact finance and lead to a devastating blow to a stakeholder. Thus, the business owner, partner, and the society are interested in bankruptcy prediction. Our stakeholder is a Taiwanese hedge fund looking to profit via short selling companies operating in Taiwan. Our business problem is a classification problem to predict if a company will go bankrupt in order to advise our stakeholder on whether or not they should decide to take a short position in a company. A short position is borrowing company stock, selling it, and buying it back when it's valued less than borrow price. It is imperative that we do not mistakenly build our model to classify companies as going bankrupt as it would result in huge financial losses since the hedge fund will need to pay back the amount of stock borrowed.

Data preprocessing is a fast and efficient way for processing imbalance data, and under sampling is widely used because of its versatility. In this project, we use the dataset from the Taiwan Economic Journal on company financial statuses from 1999 to 2009 and choose the performance metrics, recall and precision , which is suitable for the unbalanced data. Then, sampling linear models, Logistic Regression.

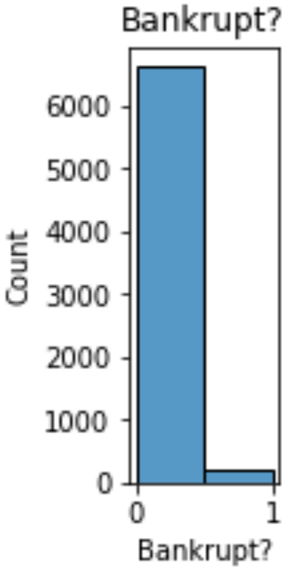
The rest of this report is as follows: research methodology, experimental results, and conclusion.

Research methodology

The process in this part consists of five parts: the first step describes data understanding; the second step describes data preparation; the third step describes evaluation indicators; the fourth step describes sub-sampling and last step describes selection model.

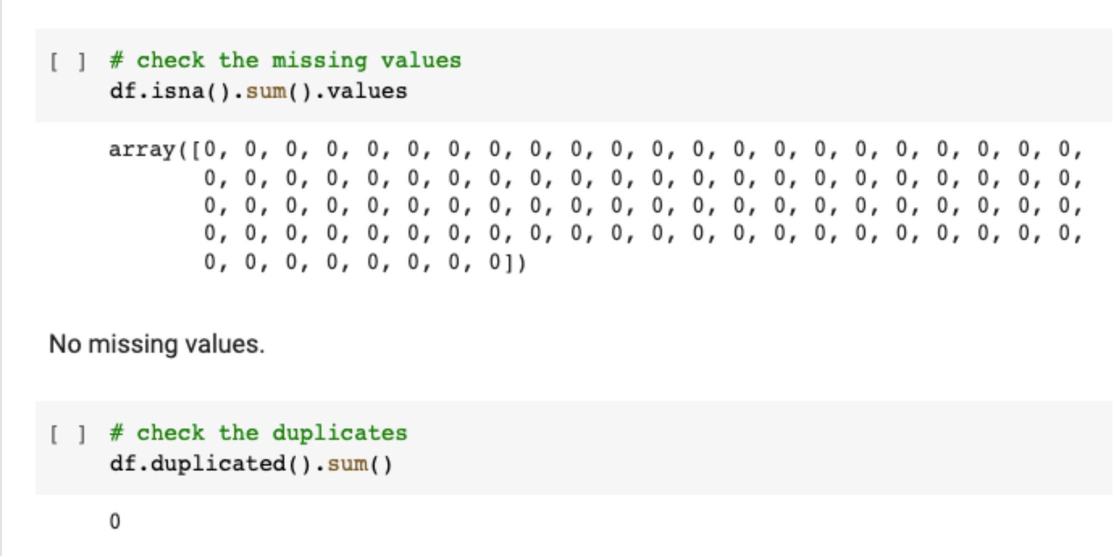
Data understanding:

The data was obtained from the Taiwan Economic Journal on company financial statuses from 1999 to 2009. It contains six thousands rows and ninety columns. The graph below states the imbalance with 97% non-bankruptcy.



Data preparation

Checking for missing values and duplicates before modeling is necessary to ensure the reliability of the data.

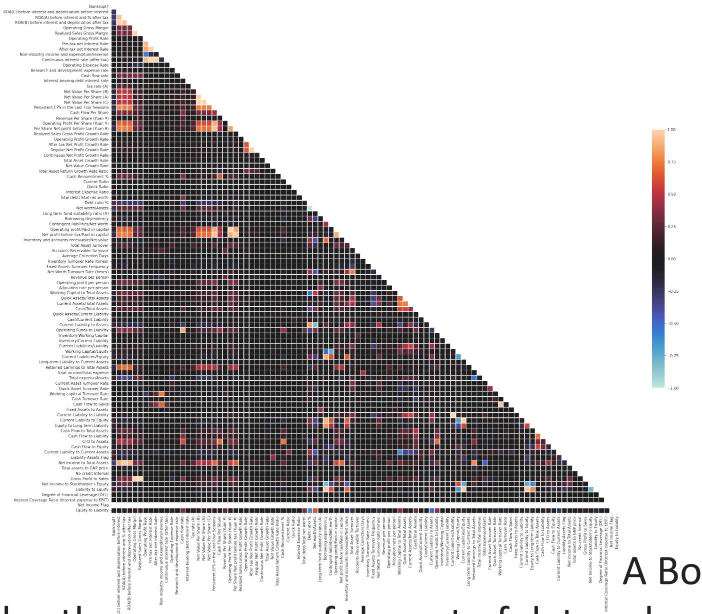


```
[ ] # check the missing values
    df.isna().sum().values

    array([0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
           0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
           0, 0, 0, 0, 0, 0, 0, 0])
```

No missing values.

```
[ ] # check the duplicates
    df.duplicated().sum()

    0
```

When plotting the correlation of all variables in the dataset. It can be seen that Net Income Flag is uncorrelated to any other columns.



A Box Plot is created to display the summary of the set of data values having properties like minimum, first quartile, median, third quartile and maximum.

The box plot uses the median and the lower and upper quartiles (defined as the 25th and 75th percentiles). If the lower quartile is Q1 and the upper quartile is Q3, then the difference (Q3 - Q1) is called the interquartile range or IQ.

A box plot is constructed by drawing a box between the upper and lower quartiles with a solid line drawn across the box to locate the median. The following quantities (called fences) are needed for identifying extreme values in the tails of the distribution:

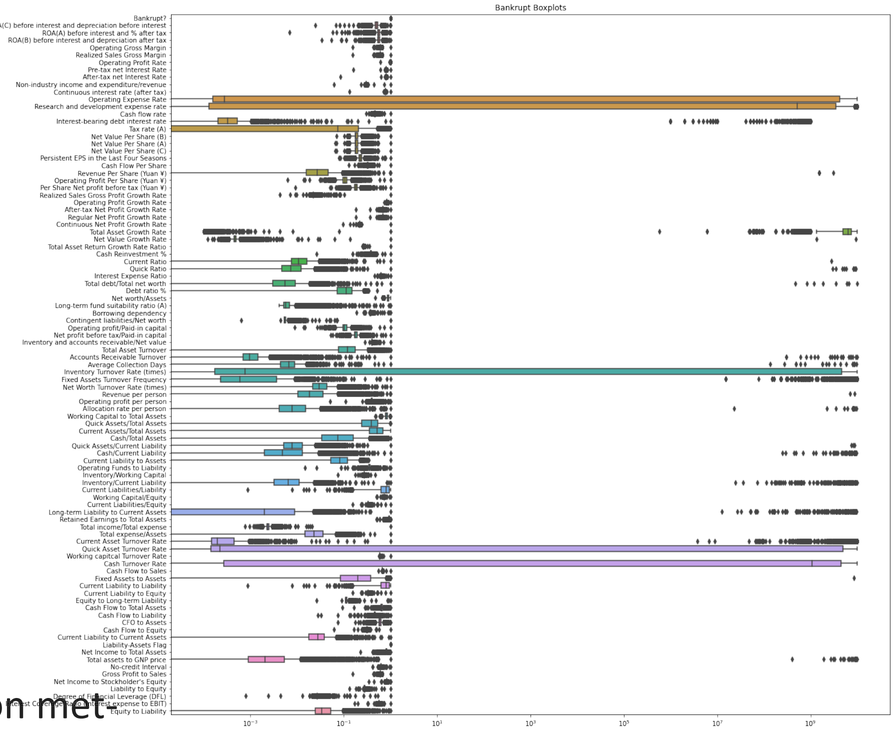
Lower inner fence: Q1 - 1.5*IQ

Upper inner fence: Q3 + 1.5*IQ

Lower outer fence: Q1 - 3*IQ

Upper outer fence: Q3 + 3*IQ

After calculating these variables, we remove the element outside the inner fence and outer fence.

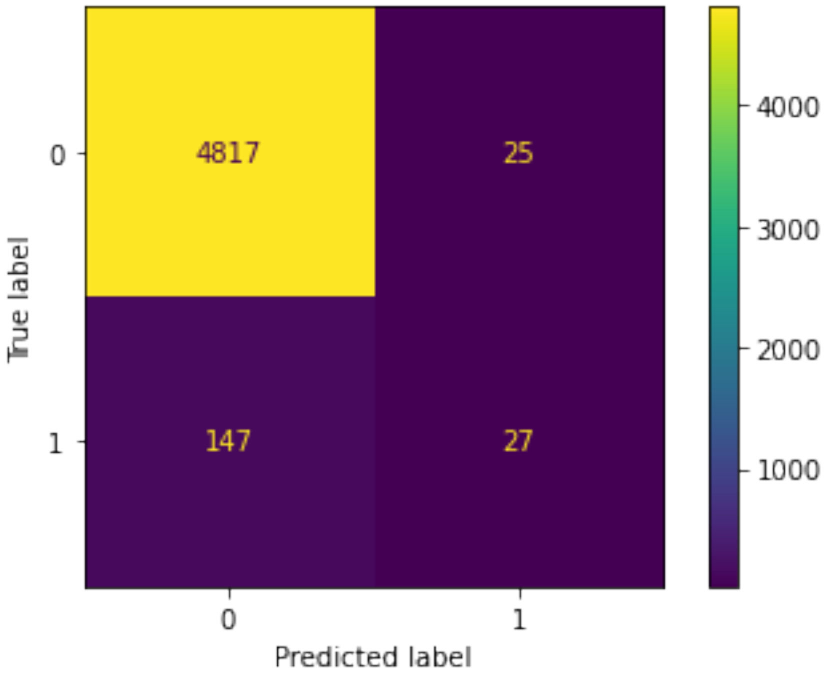


Evaluation metrics

The most widely used is the classification accuracy. The use of standard metrics in imbalanced domains can lead to sub-optimal classification models and might produce misleading conclusions since these measures are insensitive to skewed parts.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

The figure below is a confusion matrix with negative classes, ‘0’ and positive classes, ‘1’.



There is 97% of the non-bankruptcy of in this dataset. Therefore, if each instance is classified into the majority class, the accuracy will be as high as 97%, resulting in an imbalance in the traditional accuracy index. To avoid this imbalance inaccuracy when comparing different algorithms, choosing a reasonable evaluation index is necessary.

We aim to select the most suitable evaluation index according to the characteristics of the Taiwan bankruptcy dataset. Precision and recall can quantify the number of correct positive predictions.

Precision and recall have different focuses; therefore, precision is sensitive to data distribution but not to recall. The F1-measure, which combines these two evaluation indicators, is a widely used metric, evaluating the performance of algorithms dealing with data imbalances.

$$Precison = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FP}$$

$$F_1 = \frac{Precison*Recall}{Precison + Recall}$$

Undersampling methods

The method used in this project is to oversample the minority class. The simplest approach involves duplicating examples in the minority class, although these examples don’t add any new information to the model. Instead, new examples can be synthesized from the existing examples. This is a type of data augmentation for the minority class and is referred to as the Synthetic Minority Oversampling Technique, or SMOTE for short.

SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

T he algorithm is defined with any required hyperparameters, then we will use repeated stratified k-fold cross-validation to evaluate the model. We will use 5-fold cross-validation, meaning that 5-fold cross-validation is applied one time fitting and evaluating 5 models on the dataset.

The dataset is stratified, meaning that each fold of the cross-validation split will have the same class distribution as the original dataset. We will evaluate the model using the precision and recall. This can be optimistic for severely imbalanced datasets but will still show a relative change with better performing models.

Once fit, we can calculate and report the scores across the folds and repeats.

Prediction models

To avoid the extreme situation when the training set and test set are separated, this experiment uses repeated hierarchical 5-fold cross-validation. Each fold has approximately 600 examples; each fold is nearly 96% of the non-bankruptcy companies, and 3% is bankruptcy. 5-fold cross-validation repeats one time, which means each model fits 5 times to avoid a fluke. It is necessary to take measures to avoid artificially high scores caused by data leakage. This study ensured that only the training set information was used in the data processing by using 'pipeline' and carefully check the fit and transform process.

Conclusion

We iterated through our model and saw that logistic regression with under sampling and over sampling performed favorably for our business problem.

Further analysis we could pursue to better predict bankruptcy in Taiwan:

We can manually run Stratified K Folding with threshold implementation to fix the threshold issue.

Add data and information such as dates of bankruptcy

Continue to think of new parameters and different modeling strategies

How to run the code

In order to execute the Jupiter notebook, we need to follow these steps:

- Launch the Jupyter Notebook App.
- In the Notebook Dashboard navigate to find the notebook: clicking on its name will open it in a new browser tab.

- Click on the menu Help -> User Interface Tour for an overview of the Jupyter Notebook App user interface.
- You can run the notebook document step-by-step (one cell a time) by pressing shift + enter.
- You can run the whole notebook in a single step by clicking on the menu Cell -> Run All.
- To restart the kernel, click on the menu Kernel -> Restart. This can be useful to start over a computation from scratch.

Contribution

Haoyuan Zhu was responsible for data preparation, modeling planning and building, PPT production, and detailed report.

Akashay Rita was responsible for evaluation , deployment, and final report.

Priyanka Mulagandla was responsible for business understanding, data understanding, PPT preparation and final report.