Content

1. Introduction

Financial fraud can reduce confidence in the industry, destabilize economies, and affect people's cost of living. According to Beasley (Beasley,2010), the average share price of the first financial fraud business decreased by 16.7%, and around 47% of companies that have frauded records have been delisted. With such heavy loss, financial fraud detection (FDI) is vital.

In this specific business case, financial records of 2500 Shanghai A-share companies from 2006 and 2019 have been collected, with information such as Asset Structure, Portability, Cash Flow, Operational Ability, Audit, whether fraud and corresponding fraudulent years. To have a more accurate prediction, the target variable(Y) is the combination of the *year of fraud* and *fraud,* in another word, whether a company cheat in a given year, by building time-series models.

The financial fraud identification system has experienced the evolution process from single variable to multi-variable, from static to dynamic. The earliest research can be traced back to Fitzpatrick (Fitzpatrick F,1932), who first used the univariate model. Later, Bell and Carcello(Timothy B. Bell & Joseph V. Carcello, 2000) empirically analyzed 77 fraudulent and 305 non-fraudulent samples using a logistic regression model. However, Alden(Alden et al., 2012) found that the genetic algorithm is more effective in identifying financial fraud than the logistic regression model. And more currently, Chi-Chen Lin(Lin et al., 2015) used the Logistic model, artificial neural network, and decision tree to study 129 fraud samples and 447 non-fraud samples and found neural network is the most effective.

Through the logistic regression model overcomes the weakness of statistical models by improving its applicability, machine learning models became more popular by further avoiding the multilinearity that may be introduced by the logistic model. However, the "black box" analysis process of many machine learning models is still hard to explain in the business analytics background. Considering the benefits and weakness of these models,7 models including logistic regression, random forest, decision tree, Linear & RBF SVM, and MLP have been built for fraud prediction and comparison.

2. Data Exploration

2.1 Dataset Description

Financial records, between 2006 and 2019, of total of 2500 companies from 28 different industries have been provided. The dataset contains 17 general features as Table 1 presented, and excepting the first 5 features, the remaining features could be further divided into 13 sub-features that contain the information in a specific year, and therefore, there are 166 columns in total.

Stock id	Leverage ratio (2006-2019)	Inventory Growth Rate (2006-2019)
IPO time	Net profit ratio (2006-2019)	Growth rate of accounts receivable (2006-2019)

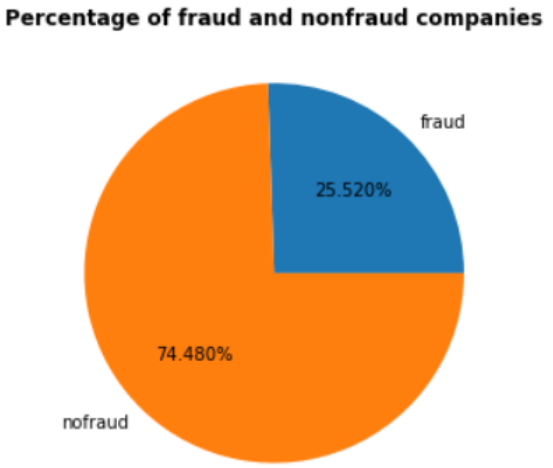
Industry	Growth rate of sales revenue (2006-2019)	Growth rate of sales management expenses (2006-2019)
Fraud	Net operating cash flow / Net profit (2006-2019)	Whether the audit by big 4 (2006-2019)
Year of fraud	Turnover of total assets (2006-2019)	Time lag of annual report disclosure (2006-2019)
Asset index (2006-2019)	Change rate of receivable turnover (2006-2019)	

(Table 1. Features of dataset)

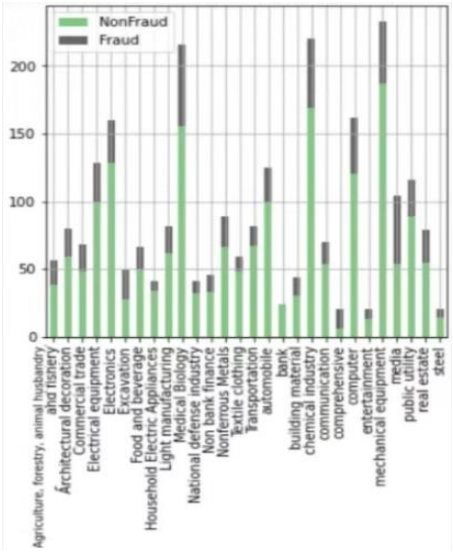
2.2 Statistic summary

2.2.1 Overall data exploration

From an overall perspective, the percentage of fraud and nonfraud is highly unbalanced, only around 25% of companies have fraud records.



(Figure 1. Distribution of fraud and nonfraud samples)



(Figure 2. Fraud distribution in the industry)

2.2.2 Industry & Fraud distribution

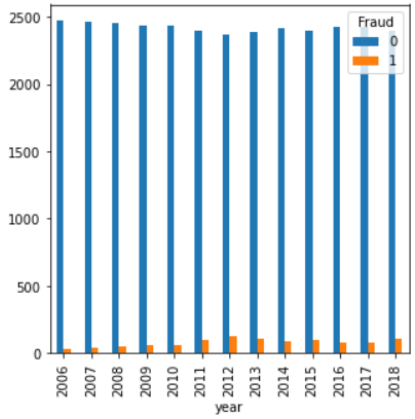
The imbalanced distribution also exists in each industry (Figure 2). Though the sum of samples size from mechanical equipment, chemical industry, and medical biology industries occupies almost 60% of the total sample, the percentage of fraud is only around 30% respectively. However, with only 20 samples, the percentage of fraud in the comprehensive industry is the highest (70%), followed by that of media, around 48%.

Bank is the most special industry as there is no fraud record. By further exploring data, it has been found that 65 columns have almost 100% missing value. With the lack of fraud samples and limited features, it could not be analyzed with other industries, and therefore, a separate One-class SVM model has been built for banks lately.

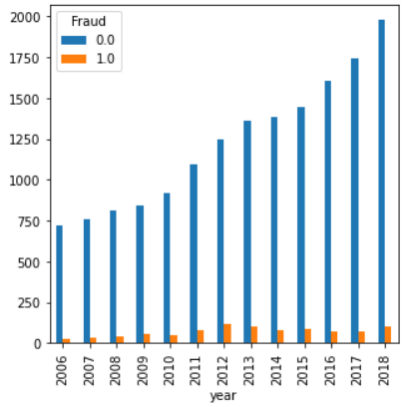
2.2.3 Year & Fraud distribution

For a better explanation, companies in each year have been treated as an independent sample, and therefore the original dataset could be converted into 32500 samples with only 17 features.

Figure 3 presented the number of fraud and nonfraud companies each year, and the number of frauds reached its peak in 2012. Since there is no fraud before IPO time, figure 4 further presented the data after filtering. There is an obvious increasing trend in the number of samples, and it is reasonable to interpret that with the development of the economy, more and more companies want to sell their stocks for fundraising. Similar to the observation from figure 3, the number of frauds is the highest in 2012, but with a higher percentage. It may relate to the financial inspection policy issued by The China Securities Regulatory Commission in 2012. Since the 2008 financial crisis, the performance of many companies has experienced a decline. Under the relatively lenient policy, fraud may not be revealed until the issue of notice on the special inspection of 2012 annual financial reports, and that is why the number of frauds increased significantly.



(Figure 3. Fraud distribution before IPO filtering)



(Figure 4. Fraud distribution after IPO filtering)

Overall, the fraud and nonfraud samples are highly unbalanced in each industry. There is no direct positive relationship between sample size and fraud percentage. In addition, after filtering data after IPO, it has been found that the number of IPO companies is increasing each year, and the number of fraud companies is highest in 2012, which may because of the issue of financial inspection notice issued by The China Securities Regulatory Commission.

2.3 Preprocessing process

The whole preprocessing process could be divided into 4 parts: data split, missing value filling, oversampling, and feature selection.

2.3.1 Data Split

To predict fraud in a given year, the dataset has been split into train, valid, and testing datasets by time sequence. Since data before 2012 have too much missing value, and long intervals may lead to bias, this report only used data from 2012 to 2019 with 5-year intervals. Table 2 presented the details of the data split.

Train dataset	X: index from 2012-2016 y: fraud in 2017
Valid dataset	X: index from 2013-2017 Y: fraud in 2018
Test dataset	X: index from 2014-2018 Y fraud in 2019

(Table 2. Dataset split)

2.3.2 Missing Value Imputation

Missing value imputation has a huge influence on model training. Simple deleting data may lose important features or samples with such a limited sample size. In this case, refilling has been conducted after the data split to avoid information leakage.

Since the business purpose is to predict future fraud, extra noise and biased data are not desired. The missing value has been refilled by considering the differences between different industries by using the mean in each industry. Though mode may not be affected by the outliers, it has not been used because most values are in decimal, and therefore it is hard to find a mode for refilling. Missing value has also been refilled in the bank industry. Columns that have missing value of over 70% have been deleted as it is meaningless for prediction. The rest are all under 40% and have been filled by using the mean value.

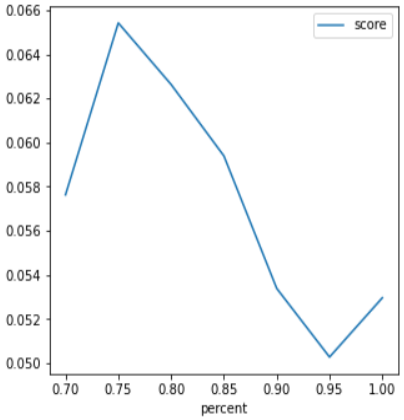
2.3.3 Oversampling

As above mentioned, positive and negative samples are highly unbalanced. Without any oversampling, the model will always predict nonfraud, with is meaningless under this business background. In addition, since the sample size is relatively small, only oversampling methods have been chosen.

Two baseline and three models have been used for data oversampling. To control the variables, oversampling is all conducted within the pipeline, and performances are all evaluated by logistic regression. As the score for nonfraud prediction are very similar and fraud prediction is the main purpose of this business case, table 3 only presented the score of fraud sample prediction. It has been found that SMOTE Tomek has the highest recall score, and Randomoversampler has the highest accuracy. Considering the penalty of miss classifying a true negative (TN) is much higher than the true positive (TP), SMOTE Tomek has been chosen for final oversampling

Method	**Precision**	**Recall**	**F1 Score**	**Accuracy**
Dummy (Baseline)	0	0	0	0.96
Logistic (Baseline)	0.11	0.01	0.02	0.95
SMOTE	0.05	0.58	0.1	0.54
SMOTE Tomek	0.05	0.64	0.1	0.54
RandomOverSampler	0.06	0.55	0.1	0.57

(Table 3. Oversampling Scores)



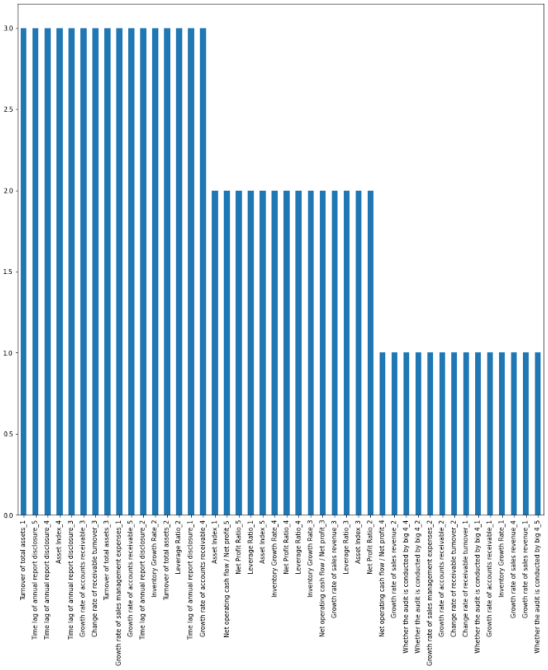
(Figure 5. SMOTE ratio score)

The oversampling ratio tuning process has also been presented in figure 5. The recall score is highest when the ratio is equal to 0.75, so the hyperparameter is also fixed.

2.3.4 Feature Selection

Feature selection is an effective approach to avoiding the curse of dimensionality, feature masking and double counting. In order to avoid this, four methods have been used for features

selection, which are random forest model-based, RFE, and GenericUnivariateSelction. Before selection, the year information in each feature has been replaced by comparative years. In another word, regardless of the actual year in the real world, the index will always start from year 1 to year 5. Features that have been selected twice and above are considered important. The top 30 features have been selected as figure 6 presents.



(Figure 6. Feature appearing counts)

3. Algorithms implementation

3.1 Logistic

3.1.1 Selection reason

Logistic regression classifier is one of the simplest binary classification models. After feature scaling, the computing process is very efficient, and the result is explainable. Since the dataset is relatively small in this business case, the logistic model is suitable.

3.1.2 Parameterization

Parameterization has been presented in table 4. Class weight treating fraud and nonfraud samples unequally to prevent the prediction bias brought by an unbalanced dataset. penalty stands for the type of regularization, and C decides the regularization intensity. L1 regularization may compress the parameter to 0, but L2 regularization will only make the parameter infinitely close to 0, but never equal to 0. Both penalty and C are used to prevent overfitting and increase the generalization of the model, without much prior knowledge, C could range from 0.001 to 100 to get the best performance.

Object	Parameter setting
Class weight	Weights = np.linspace(0.05, 0.95, 20) {0: x, 1: 1.0- x} x in weights
C	np.linspace(0.001,100,10)
penalty	['l1','l2']

(Table 4. Logistic Regression parameterization)

3.2 Decision Tree

3.2.1 Selection reason

Decision Tree is another commonly used classification model. The treemap is a good approach to examining the significance of variables or a combination of variables. In this business case, by observing the features of fraud and nonfraud groups, several key take-home tips could be taken for preliminary fraud prediction.

3.2.2 Parameterization

Decision tree parameterization has been presented in table 5. Due to the weakness of the decision tree, all these parameters are set to prevent overfitting.

Object	Parameter setting
max_features	np.arange(5,20,1)
max_depth	np.arange(2,20,1)
min_samples_leaf	[2,3,5,10]

(Table 5. Decision Tree parameterization)

3.3 Random Forest

3.2.1 Selection reason

Random forest further overcomes the overfitting weakness of decision tree, by using the bagging and ensembled method. It is also suitable for large dataset processing in the further model application, as it can handle missing value.

3.2.2 Parameterization

Similar to the parameterization of the decision tree, that of random forest adds a new parameter, which is the number of estimators, which defines the tree number. The parameter has been set in a relatively small range, as it cost at least one hour to get results when the upper limit is 50 during tunning process.

Object	Parameter setting
n_estimators	np.arange(5,20,1)
max_features	np.arange(5,20,1)
max_depth	np.arange(2,10,1)

(Table 6. Random Forest parameterization)

3.4 Linear SVM & RBF SVM

3.2.1 Selection reason

Since linear and RBF SVM use similar algorithms, they will be discussed together. Linear SVM is based on the assumption that the dataset is linearly separatable, and the object is to find the best hyperplane with the highest robustness that can best separate data from different classes. However, since no prior knowledge about whether the dataset is linearly separable, RBG SVM has also been selected by using the Gaussian kernel to project the dataset to a higher dimensional space to find the hyperplane.

3.2.2 Parameterization

With the similarity between linear SVC and logistic regression, the parameters remained same. For RBF SVM, gamma can be seen as the inverse of the radius of influence of samples selected by the model as support vectors, and it defines the impact size of a single training sample; the smaller the value, the greater the impact, and the larger the value, the smaller the impact.

Object	Model	Parameter setting
class_weight	Linear SVM	Weights = np.linspace(0.05, 0.95, 20) {0: x, 1: 1.0- x} x in weights
C		np.linspace(0.001,100,10)
penalty		['l1','l2']
C	RBF SVM	np.linspace(0.001,100,10)
gamma		[0.000001, 0.00001,0.00005,0.00008,0.0001]

(Table 7. SVM parameterization)

3.6 MLP

3.2.1 Selection reason

MLP is one of the most advanced technologies that could almost build any kind of model if required. As the introduction mentioned, it performed best in several researches, so worth to try.

3.2.2 Parameterization

There are 4 parameters in MLP. For different activations, elu and selu further prevent the death of Neurons in relu since there are no saturation points. Adam is the current best optimizer by combining momentum and RMSprop based on gradient descent. To prevent no convergence, alpha also ranges from 0.001 to 100.

Object	Parameter setting
hidden_layer_sizes	[(i*5,)*i for i in range(5,10)]
activation	['relu','elu','selu']
solver	['sgd','adam']
alpha	np.linspace(0.001,100,10)

(Table 8. MLP parameterization)

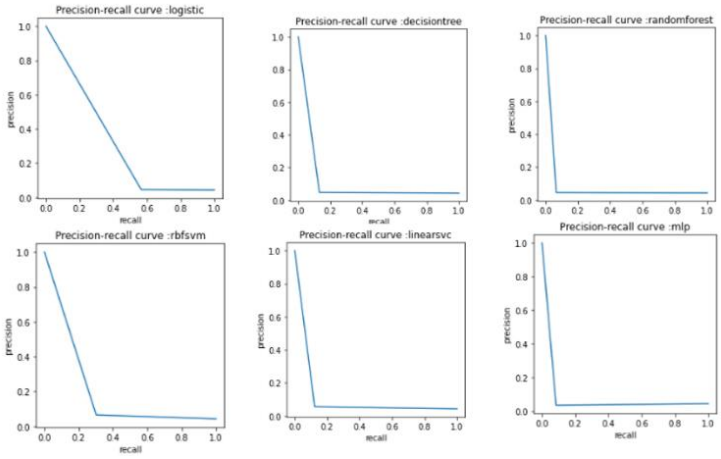
3.7 Evaluation strategy

Under this specific business case, the main purpose is to predict the fraud of companies in the coming year. Since the failure of FDI always comes with heavy loss, in another word, miss classification of fraud samples is much more expensive than that of nonfraud, recall score is the most important evaluation criterion. Precision is another metric worth paying attention to. In this case, the combination of recall and precision could be checked by the area under the precision and recall curve (AUPRC). The larger space the better performance. The final evaluation metric is robustness. The model should have similar performance on the training and testing dataset, which means that the model is well generalized.

4. Result and Analysis

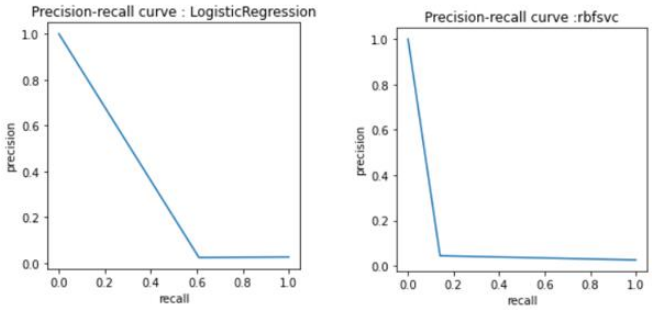
4.1 Result present

Based on the evaluation strategy mentioned above, the performances of each model have been presented below.



(Figure 7. PRC of different models)

It is obvious that the area under the curve is the largest when using logistics and RBF SVM. The recall score is around 0.58 and 0.3 respectively, and precision score are similar. The robustness of these two models has also been checked on the testing dataset and have presented in figure 8. By comparing, it has been found that the recall of logistic regression on the testing dataset has increased from 0.58 to 0.6, while that of RBF SVM decreased by 0.16. Therefore, logistic regression has been selected as it has the highest fraud prediction accuracy and best robustness.



(Figure 8. PCR on testing dataset)

4.2 Business understanding and analyzing

Recall the selected figures in figure 6, it has been found that features with larger year suffixes such as 4,5 appear are more than 1 and 2. It means that the financial index of companies that is closer to the year to predict is more important, companies that have fraud last year are highly suspected to fraud again.

Figure 9 presents the number counts of general selected features. It shows that the delay of annual reports disclosure is most important. Annual report usually contains financial information such as balance sheet, income statement, cash flow statement, and comments of auditors, and these are the indexes that most directly reflect the operation of a company. So the delay of disclosure reflects some financial problem of company. The combination of asset index and leverage ratio compares the overall debt load of a company with the assets index.

The higher the leverage ratio, the larger the risk. While the first three features mainly focus on the risk of a company, the rest of the features pay more attention to profitability such as the growth rate of inventory and sales revenue.

Time lag of annual report disclosure	5
Asset Index	4
Leverage Ratio	4
Growth rate of accounts receivable	3
Inventory Growth Rate	3
Net Profit Ratio	3
Turnover of total assets	3
Net operating cash flow / Net profit	2
Change rate of receivable turnover	1
Growth rate of sales management expenses	1
Growth rate of sales revenue	1

(Figure 9. Number of selected feature counts)

Overall, by analyzing the selected features, it shows that risk and profitability of a company are the most important criteria for fraud prediction, and companies fraud last year are very likely to fraud again.

5. Conclusion

In this report, a logistic regression and One-class SVM models have been built for company fraud prediction by analyzing 2500 Shanghai A-share companies' financial indexes from 2006 and 2019. The whole analyzing process could be divided into 4 parts: data cleaning, oversampling, model training, and model selection. Among 6 models, logistic regression has the highest recall score and most stable performance.

5.1 Potential Contribute

Compared with simply predicting the possibility of fraud, this model is more advanced and accurate by pointing out the specific fraud year. With high robustness, the model is simply applicable. In addition, by separately built one-class SVM model, prediction result is more informative then simply deleting the information.

5.1 Limitation

However, there are also several limitations. First, the dataset is relatively small, so the performance is not satisfying. Therefore, it may need reparameterization for a larger dataset in future applications. Secondly, due to the limited space, model stacking has not been applied. By combining model, the prediction accuracy may could be improved.

5.3 Future implication

In terms of the limitation mentioned above, future research can collect more samples to improve prediction accuracy. Also, stacking could be applied based on current single logistic model. It can make up the weakness of single models and improve the overall detection rate.

6. References

Alden, M., Bryan, D., & Tripathy, A. (2012). *Detection of Financial Statement Fraud Using Evolutionary Algorithms*.

Altman, E. I. (1968). The Journal of FINANCE THE PREDICTION OF CORPORATE BANKRUPTCY. *Journal of Finance*.

Beasley，M. S.，Carcello，J. V.，Hermanson，D. R. and Neal，T. L.，2010，*Fraudulent financial reporting: 1998-2007: An analysis of US public companies* Committee of Sponsoring Organizations of the Treadway Commission.

Calderon, T. G., & Green, B. P. (1994). Signaling fraud by using analytical procedures. *Ohio CPA Journal*, 53, 27-27.

Csrc.gov.cn. (2012). Notice on doing a good job in the special inspection of the 2012 annual financial reports of companies with initial public offerings of shares http://www.csrc.gov.cn/csrc/c101819/c1499917/content.shtml

Fitzpatrick, F. (1932) A Comparison of Ratios of Successful Industrial Enterprises with Those of Failed Firm. *Certified Public Accountant*, 6, 727-731

Lin, C. C., Chiu, A. A., Huang, S. Y., & Yen, D. C. (2015). Detecting the financial statement fraud: The analysis of the differences between data mining techniques and experts' judgments. *Knowledge-Based Systems*, *89*, 459–470. https://doi.org/10.1016/j.knosys.2015.08.011

Timothy B. Bell, & Joseph V. Carcello. (2000). A Decision Aid for Assessing the Likelihood of Fraudulent Financial Reporting. *A Journal of Practice & Theory*, 169–184.