1. Introduction

1.1 Main problem and background

Fraudulent companies tend to deceive the public and government by using fake information. For the long term, the relationship between public and industry will be damaged by fraud behaviors, resulting in financial loss, money investment decrease and damage to international and economic reputation. Thus, it is essential and necessary to detect company fraud behavior in time to reduce and mitigate negative impact.

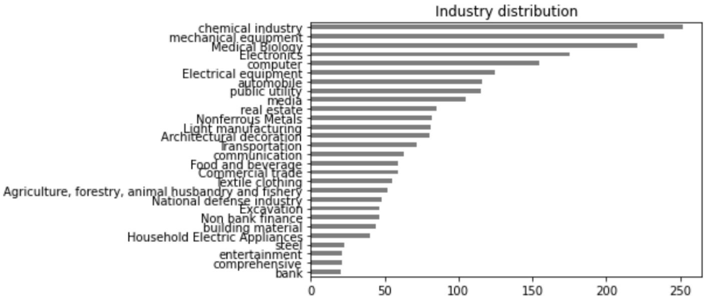
There are many types of fraud: external fraud, insurance fraud, tax fraud, billing fraud, accounting fraud and identity fraud. According to case context, fraud type in this paper belongs to billing fraud and accounting fraud. Billing fraud means that companies use false reporting of expenses for obtaining financial gain. Accounting fraud indicates that companies utilize false reporting of income statement, profits and expenses to deceive investors. For most situations, fraudulent companies are divided into two groups. One is company purposely assign illegal task to staff and make up fake report to get value (actively fraud-fraudulent financial reporting), the other is that employee intently use their positions to obtain benefits from company and company is regarded as financial fraud (passively fraud-misappropriation of assets). For this article, fraudulent companies is actively fraud group.

Traditionally, external auditors have responsibility to detect material misstatements instead of fraudulent firms. In addition, the auditor of financial statements must obtain reasonable assurance that the statements are free of material misstatements, whether caused by error or fraud. Thus, machine learning is an effective method to speed up the process of predicting fraudulent companies. Currently, the main problem is how to predicting next year’s fraudulent companies. Thus, option 2 is chosen that predict whether a company cheat in a given year, based on previous three years data.

1.2 Benefits and drawbacks of current solution

Benefits
1. Evidence-based selection. For imputers, KNN imputer, simple imputer and iterative imputer are compared and best performance imputer is concluded. Using Permutation_importance and RFE to do feature importance. Thus, overlapped features are most important.
2. Time lag reduction and efficiency process for predicting fraudulent behavior. Regulators could get fraudulent company candidates which narrow down the range of fraudulent behavior detection. To some extent, the distribution of management resources and workforce could optimized.
3. Within given dataset, linear SVC as best model is re-evaluated and has good performance.
Drawbacks
1. Only previous three years data are considered as input data
2. When fill missing value, KNN imputer takes charge and not according to each column’s meaning (context).
3. Output is limited to the input provided and the original dataset is limited.
4. Neural network is not included, which could learn new pattern of data.

2. Data Exploration



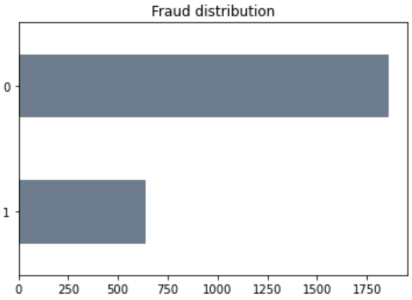


Figure 1 (top left): Industry types distribution of the dataset.

Figure 2 (top right): Fraudulent and not fraudulent company amount distribution .

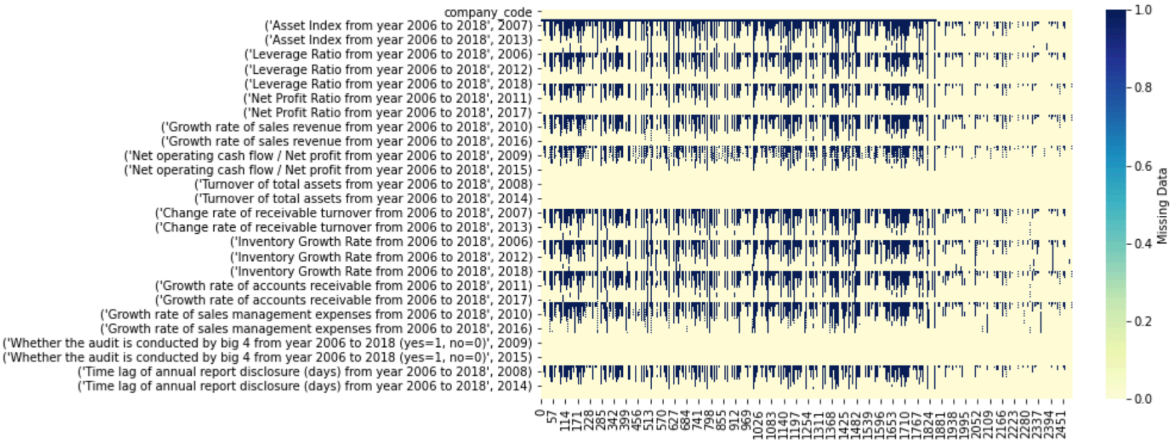


Figure 3: Missing data distribution

Manufacturing industry occupies most part of dataset, while bank and finance industry is minority. The number of not fraudulent companies is much higher than that of fraudulent companies. Ratio of fraudulent cases is 0.2552 and ratio of non-fraudulent cases is 0.7448, indicating this is a case of class imbalance problem. Through observing missing value distribution (figure 3), imputer is necessary to fill missing value.

For preprocessing process, there are main steps. Firstly, I rename the unnamed columns as company_code. Secondly, I split all time period of the same feature as one dataset. For example, asset dataset contains all time period of asset and extract 2013, 2014, 2015 data by using iloc function for x train dataset. There are 14 divided features relevant dataset for train, validation and test dataset merge. For categorical feature ‘big_4’, replace its ‘--' value as nan and change its datatype to float. Adding relevant years and three different situations (no, yes, unknown) as new feature names. Then, one-hot encoder is used for transforming ‘big_4’ to numeric array.

Besides, Year_of_Fraud is changed to time in seconds passed since epoch in local time for calculating. For target variable preparation, I define a function to replace target years (e.g. 2016) with 1, and other years return 0. Then, add corresponding column name to the dataset. For dataset selection, 2013, 2014 and 2015 data as x_train dataset, companies which fraud in 2016 as y_train. 2014, 2015 and 2016 data as x_validation dataset, companies which fraud in 2017 as y_validation. 2015, 2016 and 2017 data as x_test dataset, companies which fraud in 2018 as y_test.

For imputer selection, company_code is dropped and all data is transferred as float datatype for calculating. There are three imputers: KNN Imputer, IterativeImputer and SimpleImputer. KNN imputer is compared against with other two imputers to conclude which imputer has highest recall score. As a result, knn imputer is selected to fill null value. Because it can learn non-linear decision boundaries when used for classification problem and mean imputation that it tends to produce bias estimates for some parameters, particularly for the variance.

After KNN imputation, this dataset is not balanced through observation of target variable (y). Smote (over-sampling) method deal with this issue for train and validation datasets. However, smote only works well if the minority case features are similar.If fraud is spread through the data and not distinct, using nearest neighbors to create more fraud cases, introduces noise into the data, as the nearest neighbors might not be fraud cases.

Balanced data is standardized through StandardScaler. Next step is feature importance and selection. Permutation_importance and RFE are two methods for selecting most important features to consider. See below table 1, seven overlapped features are filters for all dataset as input data. New meaningful features name replaces original feature names to know directly column meanings.

Table 1: feature selection and new defined feature name

7 overlapped important features (train)	**7 overlapped important features (validation)**	**New meaningful features name**
IPO_Time	IPO_Time	IPO_Time
('Time lag of annual report disclosure (days) from year 2006 to 2018', 2014)	('Time lag of annual report disclosure (days) from year 2006 to 2018', 2015)	'Time lag_middle_year'
('Time lag of annual report disclosure (days) from year 2006 to 2018', 2013)	('Time lag of annual report disclosure (days) from year 2006 to 2018', 2014)	'Time lag_earlier_year'
('Leverage Ratio from year 2006 to 2018', 2013)	('Leverage Ratio from year 2006 to 2018', 2014)	'Leverage Ratio_earlier_year'
('Inventory Growth Rate from 2006 to 2018', 2015)	('Inventory Growth Rate from 2006 to 2018', 2016)	'Inventory Growth Rate_final_year'
('Growth rate of accounts receivable from 2006 to 2018', 2015)	('Growth rate of accounts receivable from 2006 to 2018', 2016)	'Growth rate of accounts receivable_final_year'
('Net Profit Ratio from year 2006 to 2018', 2013)	('Net Profit Ratio from year 2006 to 2018', 2014)	Net Profit Ratio_earlier_year'

3. Algorithms implementation

3.1 Models selection and parameterizations

Adaptive boosting classifier, dummy classifier, logistic regression, linear support vector classification (linearSVC), random forest classifier and C-Support Vector Classification (SVC-RBF) are used to be model selection. Dummy classifier is regarded as baseline for other model's performances. No parameters set in dummy classifier.

Two ensemble models are used: one is adaptive boosting classifier, the other is random forest classifier. AdaBoos could converge a strong learner with a great generalization power that not influenced by outliers and is less to overfitting. The disadvantage is that AdaBoost's sensitiveness to outliers and noisy data. Base estimator criterion, base estimator splitter, n estimators and learning rate are four parameters in AdaBoost.

Random forest classifier mitigate overfitting issue and speed up training process but need to tune lots of parameters to improve performance. There are four parameters: n_estimators, max features, max depth and criterion. N estimators and max features controls the amount of trees in the forest and amount of features to be considered separately. Max depth limits the extent of node could be splitter in the forest. Criterion determines how the impurity of a split will be measured. Through GridSearchCV, best parameters will give the best results on the hold out data.

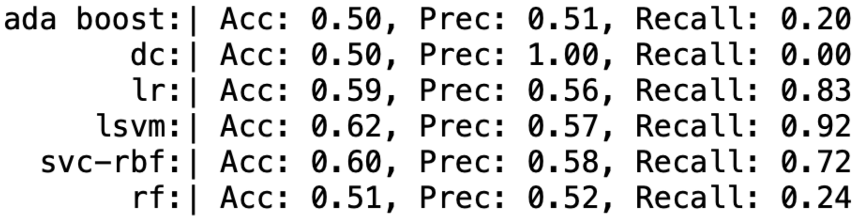
Logistic regression doesn't have assumptions about distributions of classes and constructs a linear boundary thorough data. Besides, this model is efficient at classifying unknown records. The limitation is non-linear cases can not be solved and is tough to obtain complex relationships. Penalty and C hyper-parameter which are norm of the penalty and error term for controlling error. When there are many features, L1 norm mitigates overfitting problem by generating sparse solutions. Higher C indicates more weight give to the training data, and a lower weight to the complexity penalty. I give a range of these two parameter: 'penalty':

['l1', 'l2', 'elasticnet', 'none'], 'C':[0.0001, 0.001, 0.01, 0.1, 1, 10, 100, 1000]. The best parameters are concluded from GridSearchCV.

SVC has more flexibility that all about select the right kernel and parameters. For example, RBF can capture much more complex relationships between datapoints without having to perform difficult transformations on my own. But SVC 's overfitting is quite difficult to detect or fix at times. Three parameters of LinearSVC are set: penalty, C and loss. Loss function measures the quality of solution, while penalty function imposes some constraints on solution. Two parameters of SVC-RBF is C and kernel. Through GridSearchCV, best parameters of SVC is concluded. During the process of observation, add larger number or smaller number to find most suitable parameter. For example, if a range [1, 10, 100] is given, and result shows best parameter is 100, the maximum number should be revised.

3.2 Evaluation strategy

Mean absolute error (MAE) and mean squared error (MSE) are not considered as part of evolution strategy because they are for regression models evaluation. Accuracy, precision and recall score are used to evaluate each model's performances by using train and validation datasets. Accuracy score is the most important, which is the ratio of all correctly predicted cases whether positive or negative and all cases in the data. The second put most emphasis on is recall score. For the reason that number decrease of fraudulent companies regard as non-fraudulent is more essential than predict non-fraudulent companies as fraudulent. These two scores comparison is the main condition to select best model. LinearSVC (the forth one) is selected to be final models, resulting from its top recall score and relatively high accuracy score.



```
ada boost:| Acc: 0.50, Prec: 0.51, Recall: 0.20
       dc:| Acc: 0.50, Prec: 1.00, Recall: 0.00
       lr:| Acc: 0.59, Prec: 0.56, Recall: 0.83
     lsvm:| Acc: 0.62, Prec: 0.57, Recall: 0.92
  svc-rbf:| Acc: 0.60, Prec: 0.58, Recall: 0.72
       rf:| Acc: 0.51, Prec: 0.52, Recall: 0.24
```

Figure 4: Accuracy, precision and recall scores of 6 models

Receiver operating characteristic (ROC) is used for probabilistic models which predict the probabilities of the class. ROC score of three best-performance models are calculated to compare by using each models' best parameters. True Positive Rate (TPR) is considered as part of evaluation strategy. Considering higher value of TPR would mean that the value of false negative is very low which would mean almost all positives are predicted correctly. However, ROC curves are appropriate when the observations are balanced between each class, whereas precision-recall curves are appropriate for imbalanced datasets. According to context, confusion matrix score is focused as final model selection.

4. Results presented and analysis

4.1 Performance evaluation

Table 2: Confusion matrix scores of three best-performance models performance

Final Model	Evaluation
LinearSVC	Accuracy: 0.36, Precision: 0.05, Recall: 0.85
Logistic regression	Accuracy: 0.30, Precision: 0.04, Recall: 0.90
SVC-RBF	Accuracy: 0.42, Precision: 0.04, Recall: 0.66

Using test dataset to obtain the performance scores. LinearSVC is selected to be final model, resulting from its top recall score and relatively high accuracy score. Other two models are compared. Logistic regression and SVC-RBF is the second and third best model separately. The confusion matrix score is shown above (Table 2). LinearSVC has relatively low recall score but high accuracy. By contrast, logistic regression has higher recall score but lower accuracy score. SVC-RBF has best accuracy score. Furthermore, best model is re-evaluated by ROC score to reconfirm it has good performance.

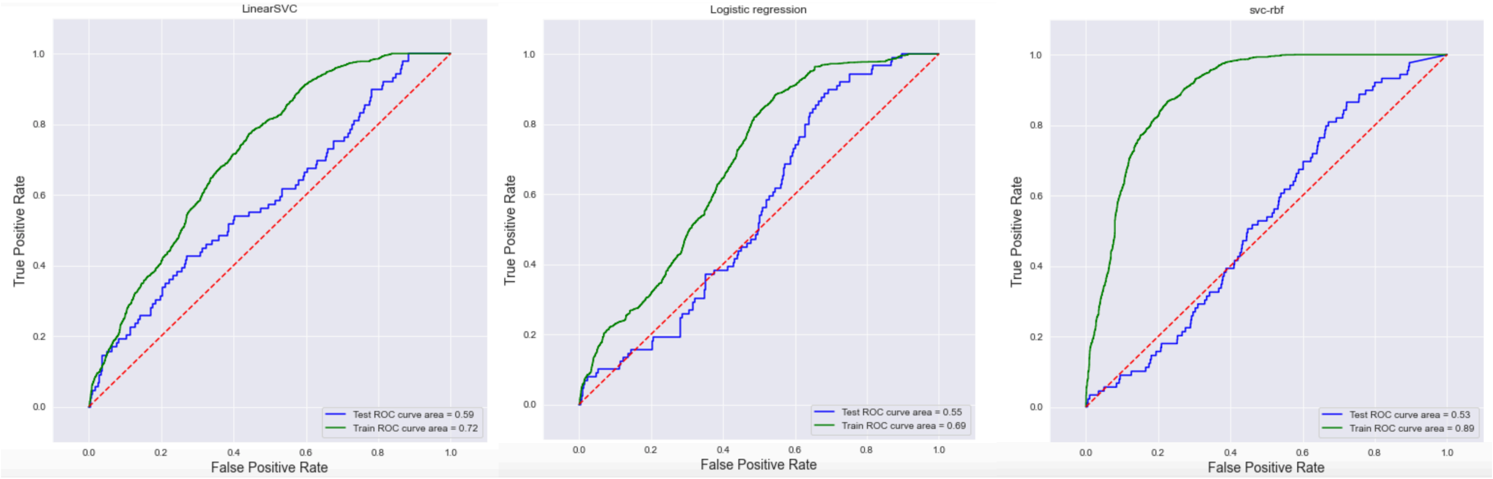


Figure 5: ROC score between linear SVC, logistic regression and SVC (RBF)

The closer the curve follows the TPR axis and then the top border of the ROC space, the more accurate the test. Meanwhile, true positive rate (TPR) is high and the false positive rate is low. See above figure 1, linear SVM occupies more space under the curve, indicating that more true positives while minimizing the percent of false positives. Accordingly, linear SVM does have better ability to discriminate.

4.2 Recommendation: for how to prevent and detect fraud behavior

Through feature importance selection, these 7 important features are filled. Each of them represents various meanings for company. Time lag of annual report disclosure could be days between fiscal year-end and annual earnings release date or days between annual earnings release date and financial statements disclosure date. Thus, there is no available financial reports in time to inform decision making, which reflect their management ability. Leverage ratio attempts to highlight cash flow relative to interest owed on long-term liabilities. Plus, it indicates the level of debt incurred by a business entity against several other accounts in its cash flow statement and income statement.

Inventory Growth Rate shows the extent of a balance between having enough inventory on hand and not having to reorder too frequently. Growth rate of accounts receivable measures a company's effectiveness in collecting its receivables by clients. If a company that is able to run fine with little cash may have very small fixed costs or a low amount of debt in its capital structure. Net Profit Ratio depicts the relationship between the net profit after taxes and net sales taking place in a business.

For financial statement fraud (passively fraud), firms are suggested to do internal audit regularly and do surprise audits. More specifically, accounting about sales and expenditure and working process could be conducted for fraud detection. It is important to pay attention to partners and clients identity. Not qualified corporation may use fake information to make transactions with firm, which increase the possibility of account receivable. Moreover, corporation usually not have much awareness about internal employees. Asset misappropriation is another reason that deceptively increase inventory growth rate. Thus, different financial tasks should be assigned to different employees who are in charge of cash-handling.

5. Conclusions

In order to mitigate the impact of fraud behavior, how to detect company fraud behavior is current concern for regulators. Accordingly, machine learning method is used to create model-based predictor to speed up the process of fraud behavior detection. In addition, 7 financial-related features that need to pay most attention are extracted from 162 columns. Using previous three years data as input could predict next year fraudulent companies. The best model is linear SVM due to its relatively good confusion matrix score and ROC score.

For passively fraud firms, scheduling regular audit need to paid most attention for checking whether there is misappropriation of assets. Not fully trust employees and assign financial tasks to various employees is another way to reduce the risk of internal fraud. For actively fraud firms, model-based predictor could detect their irregular behavior more effectively, compared with traditional methods.

However, if fraudulent companies produce a new pattern to cheat and make up fake information, it is hard to detect these companies. Moreover, neural network is not considered as part of model selection, and the output that is limited to the input provided. In the future, more data should be considered like previous five years data and combined with neural network method.