Automated identification of fraudulent financial statements by analyzing data traces

Evangelos Chytis[1] Vasileios Nastos[2] Christos Gogos[2]
Angelos Dimitsas[2]

[1]Department of Accounting and Finance
University of Ioannina, Preveza, Greece

[2]Department of Informatics and Telecommunications
University of Ioannina, Arta, Greece

7th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM 2022)
September 26, 2022

Fraud definition

- A false representation by means of a statement or conduct, in order to gain a material advantage(Oxford Dictionary).
- An international act by one or more individuals among management, those charged with governance, employees, or third parties, involing the use of deception to obtain an unjust or illegal advantage(IAASB of IFAC).

Fraud expose examples

- ENRON scandal in USA(2001)
- WorldCom-11 billion(2002)
- Tyco International-150 million(2002)
- Satyam-1.5 billion(2009)
- Folli-Follie Greece-(2018)

Dataset

Data were collected from the certified financial database ICAP Data-Prisma, the Athens Exchange (Hellex SA) and the Hellenic Capital Market Commission (HCMC).

Industrial sector	ICB-Code	Firms	%	Non-FFS	FFS FYs		%	Non-FFS FYs	FFS FYs
Chemicals	13	3	2.4%	1	2	40	4.2%	34	6
Basic Resources & Construction and Materials	17 & 23	17	13.7%	8	9	162	17.2%	128	34
Industrial Goods & Services	27	42	33.9%	31	11	251	26.6%	209	42
Food, Beverage & Tobacco	35	11	8.9%	6	5	101	10.7%	82	19
Personal Care, Drug & Grocery Stores	37	10	8.1%	3	7	76	8.1%	50	26
Health Care	45	6	4.8%	3	3	62	6.6%	46	16
Retail	53	13	10.5%	9	4	45	4.8%	36	9
Media	55	8	6.5%	4	4	74	7.8%	59	15
Travel and Leisure	57	8	6.5%	5	3	62	6.6%	46	16
Utilities	75	2	1.6%	2	0	22	2.3%	22	0
Technology	95	4	3.2%	1	3	48	5.1%	25	23
Total		124	100%	73	51	943	100%	737	206
								78%	22%

- Samples: 943
- Country: Greece
- Companies: 124(943 firm-years)
- Examination Period: 2005-2018

Excluded companies

Initially, 148 company groups were identified.

Excluded ICB Code	Excluded Sector	Firms	Firm-Years
85	Insurance	1	10
86	Real estate	3	32
83 & 87	Financial services	5	22
88	Holding	15	21
		24	85

Risk indicators

- A set of variables that correspond to financial figures and ratios were selected as representative of the status of each company.(29 total).
- Intra-group accounting transactions are excluded.
- 7 quantitive variables from financial statements(F1-F7).
- (F1-F4) variables refers to the degree of asset capitalization.
- (F5-F7) variables measure the company's ability to generate cash-flow.
- 22 Financial ratios(R1-R22).
- 1 target value(Fraud) containing two class(YES-NO).

Features

Assets (intensive)
F1
F2
F3
F4
Cash flow activities
F5
F6
F7
Profitability Ratios
R1
R2
R3
R4
R5
R22

Capital Structure Ratios
R6
R7
R8
R9
R10
R11
Liquidity Ratios
R12
R13
Activity Ratios
R14
R15
R16
R17
R18
R19
R20
R21

Workflow

Preprocessing

1. Missing values handling.
2. Values normalization.
3. Feature selection.
4. Balancing dataset.

Learning-Assessing

1. Classification
2. Model performance evaluation

Preprocessing 1/5

Missing values

1. Mean: Each missing value replaced by the mean value of the corresponding column.
2. Median: Each missing value replaced by with the median value of the corresponding column.
3. k-NN: Each missing value replaced by the mean value of the k-nearest neighbors discovered in the training set(k=10).

Preprocessing 2/5

Scaling

1. Normalization: Each value scales to the range 0 and 1.
2. Z-score normalization: Each value scales to the range -1 and 1.
3. Robust: Each value is squeezed to the interquartile range.

Preprocessing 3/5

Feature selection: In order to reduce the number of independent variables five different methods were used:

1. Variance Threshold: Elimination of the features with variance rate lower than the predetermined value.
2. ReliefF: Identification of the most relevant problem features.
3. RFE(Recursive Feature Elimination): Examination of features recursively in order to reshape the features set in to the desirable size.
4. FFS(Forward Feature Selection): Based on an estimator's performance, each feature is added to a list of features, that is initially empty, until a threshold is reached in the list length.
5. BFE(Backward Feature Elimination): Based on an estimator's perfarmance and a given numeric theshold, a number of features are eliminated from the feature list.

Preprocessing 4/5

Dimensionality reduction:

- The features that are most pertinent to the problem are used to generate a new feature list using a statistical technique. The PCA (Principal Component Analysis) approach is the most common method to accomplish that.
- Better results can be obtained by the feature selection approach in our problem.

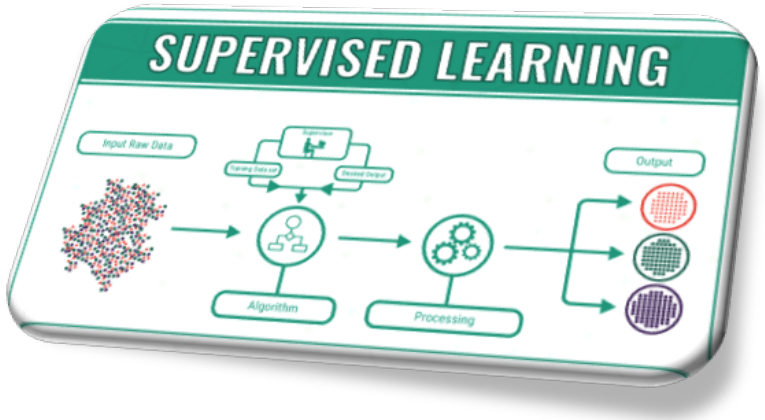
Preprocessing 5/5

Balancing Data: When the input data are unbalanced, machine learning typically performs poorly. Three balancing methods were used:

- Oversampling: Randomly selected instances of the minority and re-produce the training dataset in order to balance the data.
- Undersampling: Elimanates instances of the majority class in the training set.
- SMOTE: Based on already existing data instances, produces new instances, until it reaches the size of the majority class, the minority class is augmented with fresh synthetic data.

Classification

- Decision trees
- Random Forest
- K-nearest neighbors
- Support Vector Machine
- Naive Bayes
- Ada Boost



Cross validation

10 Fold cross validation:

- In order to estimate the skill of our machine learning workflow on new data and lower the bias a 10 fold cross validation procedure was applied to our data.
- A splitting procedure applied to the data and separates them in to 10 folds.
- 9 folds were used for training and 1 fold for test with specific test size.
- All combinations of preprocessing steps and classification algorithms are executed per fold.
- The procedure repeats 10 times, so that every one of the 10 folds will serve as a test set.
- 10800 models are produced from the procedure.

Evaluation 1/5

The following metrics are used to evaluate each unique model of the k-Fold procedure:

- Confusion Matrix
- Accuracy Score
- Recall Score
- Precision Score
- F1-Score
- Cohen's Kappa Score

Evaluation 2/5

Confusion Matrix A numerical tableau showing the behavior of the model.

- TP: Number of fraud predicted samples that are actual fraud.
- TN: Number of non-fraud predicted samples that are actual non-fraud.
- FP: Number of fraud predicted samples that are actual non-fraud.
- FN: Number of non-fraud predicted samples that are actual fraud.

	Predicted (Positive=non-fraud)	Predicted (Negative=fraud)
Actual (Positive=non-fraud)	TN	FP
Actual (Negative=fraud)	FN	TP

Evaluation 3/5

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - Measure = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{4}$$

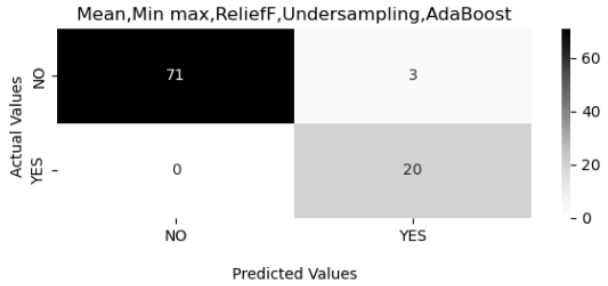
$$\kappa = \frac{2 \times (TP \times TN - FN \times FP)}{(TP + FP) \times (FP + TN) + (TP + FN) \times (FN + TN)} \tag{5}$$

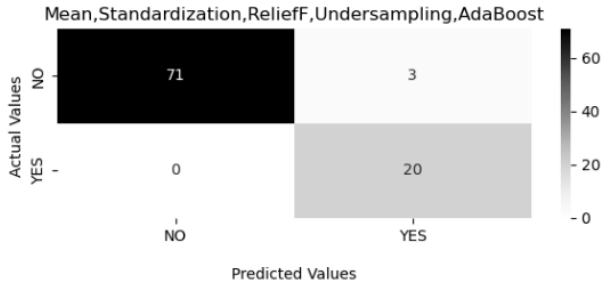
Evaluation 4/5

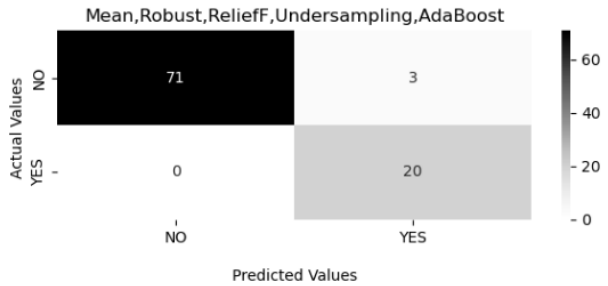
Id	Missing Values	Scaling	Feature Selection	Balancing	Classifier	Accuracy	Recall	Precision	F-measure	Cohen's kappa
1	Mean	Min max	Relief	Undersampling	AdaBoost	0.9681	1.0000	0.8696	0.9302	0.9097
2	Mean	Standardization	Relief	Undersampling	AdaBoost	0.9681	1.0000	0.8696	0.9302	0.9097
3	Mean	Robust	Relief	Undersampling	AdaBoost	0.9681	0.9500	0.8696	0.9302	0.9097
4	Mean	Standardization	Relief	SMOTE	AdaBoost	0.9681	0.9500	0.9048	0.9268	0.9064
5	k-NN impute	Maximum absolute	FFS	Oversampling	AdaBoost	0.9574	0.8000	1.0000	0.8889	0.8630
6	Mean	Robust	Relief	SMOTE	Random Forest	0.9574	0.8571	0.9474	0.9000	0.8731

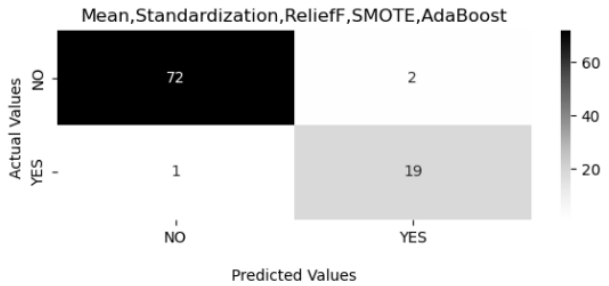
- In order to identify the best performing classifiers the pareto front of all runs was computed, against five performance metrics.
- 6 models were obtained and AdaBoost dominates in our problem as it used to five of them.

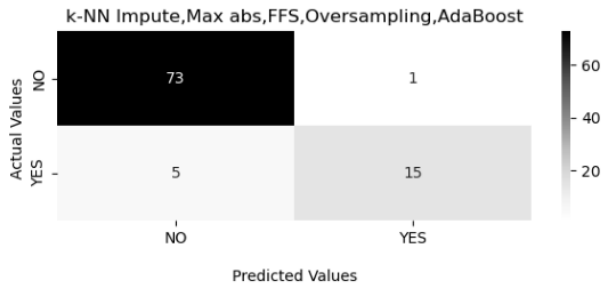
Evaluation 5/5

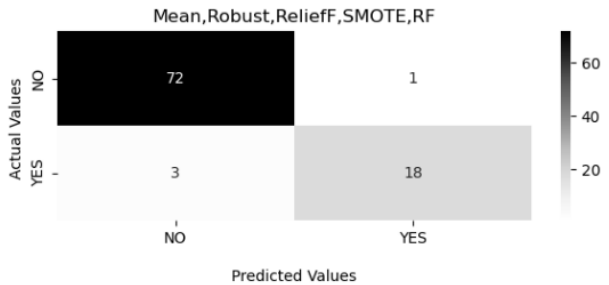












Conclusions

1. A dataset of 943 firm-years of Greek companies (2005-2018) was constructed.
2. Based on modern software tools (Python, Scikit-Learn, etc.), a machine learning infrastructure was proposed, having the potential to assist in detecting fraud in financial statements.
3. High performance scores are achieved by using state of the art machine learning algorithms.
4. AdaBoost performed really well in this problem.

The End

Questions? Comments?