FITE7410A Financial Fraud Analytics – Project Report – Group 15

Analysis and Fraud Detection of The Enron Scandal Dataset

Chung Wai Kei	**3035473488**	**Wong Ngai Sum**	**3035380875**
Lin Ho Ching Janus Lin	**3035754181**	**Tse Wai Tung Jonathan**	**3035908976**

1. Objectives

The goal of this project is to study the Enron Scandal case to train our machine learning model based on the financial and email features from the open-sourced dataset that already pre-processed from the original email data for identifying the person of interest (POIs) from the company's internal communication channels so as to detect the possibility of fraud. With the machine learning model, it could significantly avoid heavy manual work for auditing teams to discover and utilize hidden information to detect potential POI.

2. Case Description

2.1 Case Background

Enron corporation was established in 1930 and is headquartered in Houston, USA. Enron was one of the world's largest power natural gas and telecommunications companies that was mainly responsible for large-scale purchasing and trading of natural gas. In 2000, it disclosed a huge turnover of 101 billion U.S. dollars. The company has been selected by Fortune Magazine as the "Most Innovative Company in America" for six consecutive years. However, as it was later revealed, many of the profits reflected in Enron's financial statements were inflated or even fabricated out of thin air. The company kept the loss off the balance sheet through a dazzling series of financial related transactions. After years of carefully planned and institutionalized financial fraud scandals, it went bankrupt in a few weeks in 2002.

2.2 Risk and Red Flags

The Enron case shows a serious crisis in the accounting and auditing system. We summarized the following three main conditions that caused the Enron scandal:

Firstly, Black-box operations, using high risk accounting practices to transfer debts and bad debts to branch companies and overestimate future profits. Enron was using off-balance-sheet special purpose vehicles (SPVs), also known as special purposes entities (SPEs), to hide its mountains of debt and toxic assets from investors and creditors. Also, MTM (mark-to-market) accounting techniques were adopted for earnings management, that overestimate future unrealized profits, or to hide or understate future unrealized losses from soured contracts. These approaches are a kind of "accounting falsification".

Secondly, senior management of Enron had conflicts of interests. Insider trading, also known as insider trading, here specifically refers to the company's internal securities trading based on a large amount of inside information. Louis Borget and Thomas Mastroeni received internal news that enabled the company to obtain more substantial profits in oil transactions. After the incident was discovered by the auditor, Kenneth Lay, CEO of Enron, supported the auditor to conduct further investigations to "recover every penny," but he did not immediately pursue the person responsible.

Thirdly, Enron's bad culture encourages their employees to take risks. What Enron encourages is the adventurous spirit of pursuing profit at all costs, using high profits in exchange for high rewards, high bonuses, high rebates, and high options.

3. Data Description and Preprocessing

3.1 Data Source

The original dataset is an email dataset which was collected by the Federal Energy Regulatory Commission, with its latest version released in 2015. It contains ~0.6M emails organized into folders from 158 users, with most of them being senior management of Enron. However, as our aim is to study the characteristics of persons of interest (POI) and potential red flags, we picked an open source POI dataset from GitHub, which has been combined with the original dataset, to help develop fraud detection models. The persons of interest are people who were suspected to commit fraud in the Enron Scandal case. The objective of our project is to train fraud detection models that can identify POIs in the case and similar situations.

Data source: https://github.com/missmariss31/enron/blob/master/enron.csv

3.2 Exploratory Data Analysis

3.2.1 Attribute Types

Attributes of the dataset can be roughly divided into three categories: financial attributes, email attributes, and class label "POI". The details of these attributes are as follows.

Financial attributes (in US dollar)	"salary" "deferred_payments" "total_payments" "loan_advances" "bonus" "resticted_stock_deferred" "deferred_income" "total_stock_value" "expenses" "exercised_stock_options" "other" "long_term_incentive" "restricted_stock" "director_fees"
Email attributes (in # of messages, except email_address)	"to_messages" "from_messages" "email_address" "from_this_person_to_poi" "from_poi_to_this_person" "shared_receipt_with_poi" "name"
Class label (True = is POI)	"poi"

3.2.2 Missing Data

One of the challenges to produce a good fraud detection model is the large amount of absence of data. 1318 values are missing, which accounts for 41.3% of all data. In addition, the dataset is small with only 145 records. Thus, data imputation, record removal, and variable selection may be needed to reduce biases. The details of missing data are as follows.

name	0	**total_stock_value**	20	**bonus**	64	**poi**	0
salary	51	**expenses**	51	**restricted_stock**	36	**director_fees**	129
to_messages	59	**loan_advances**	142	**shared_receipt_with_poi**	59	**deferred_income**	97
deferral_payments	107	**from_messages**	59	**restricted_stock_deferred**	128	**long_term_incentive**	80
total_payments	21	**other**	53	**email_address**	0	**from_poi_to_this_person**	59

exercised_stock_options	44	**from_this_person_to_poi**	59	**Total number of missing data**	1318	**Proportion of missing data**	41.3 %

3.2.3 POIs

18 persons are identified as persons of interest, which accounts for roughly 12% of records. The class distribution is highly imbalanced as most of the records are POI, thus resampling on the training set is needed after train/test split. Also, recall and precision metrics should be adopted instead of accuracy to have an all round evaluation on the performance of fraud detection. Also, it is noteworthy that most of the POIs are senior management. There should be some big differences between senior and non-senior persons on data values, but due to the large amount of absence of data values, classification results will likely be poor. The persons of interest and their titles are as follows.

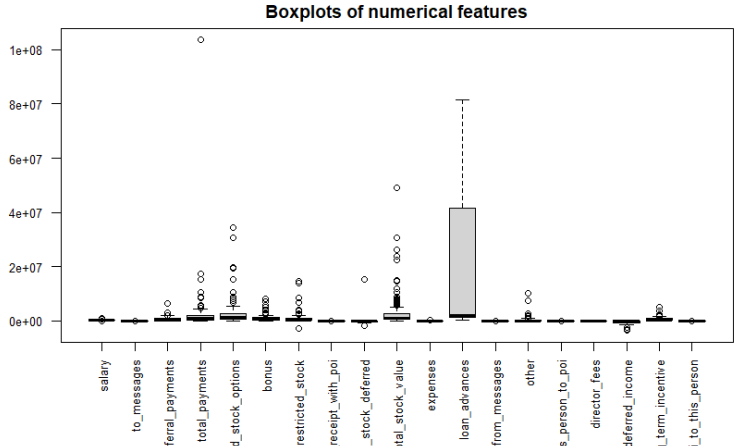
HANNON KEVIN P	Chief operating officer	DELAINEY DAVID W	Chief Executive Officer	CALGER CHRISTOPHER F	Vice president
COLWELL WESLEY	Chief Accounting Officer	LAY KENNETH L	CEO	RICE KENNETH D	Co-CEO
RIEKER PAULA H	Managing Director	BOWEN JR RAYMOND M	Chief Financial Officer	SKILLING JEFFREY K	CEO
KOPPER MICHAEL J	Executive	BELDEN TIMOTHY N	Chief Financial Officer	YEAGER F SCOTT	Senior Vice-President
SHELBY REX	Senior Vice-President	FASTOW ANDREW S	Chief Financial Officer	HIRKO JOSEPH	Co-chief Executive Officer
KOENIG MARK E	Executive Vice-President	CAUSEY RICHARD A	Chief Accounting Officer	GLISAN JR BEN F	Treasurer

3.2.4 Visualizations of data

There are a small number of outlying values from most histograms. However, the probability distributions of some variables have high variance, such as “restricted_stock_deferred”, “director_fees”, “loan_advances”. These 3 variables are more than 80% missing values. So they may not be very helpful for fraud detection. In some scenarios outliers should be excluded from model training as they are noise in the dataset, but in this case, those outlying values may exist for some benign reasons and are indications of potential fraud. Thus most records are preserved except for some records that are irrelevant or have lots of missing values, which will be elaborated later. The histograms have been shown in *Figure 1*.

3.2.5 Outlier Analysis

Further outlier analysis was attempted by plotting boxplots *(figure on the right)* of every numerical variable in the dataset. There are 19 numerical variables and so 19 boxplots were made. Similar to the above histograms, the



boxplots show the distribution of each feature, but with a representation that reveals departures from quantiles much more clearly.

No rigorous methods were used to identify outliers and outlier detection was simply based on judgment. One record has been judged to be an outlier due to its extremely large departure from other records in terms of the "total payments" value. Upon closer examination of the record, it is revealed to be the record of a POI named "Lay Kenneth L". Due to the scarcity of POI records and the unpredictable nature of frauds, it has been decided that this record should be kept for further analysis.

In conclusion, it was decided that outliers should be kept due to the small dataset size and because potential fraud may exhibit unusual behavior akin to outlier observations. It was also decided that imputation should be performed due to the abundance of missing data.

3.3 Data Preprocessing

3.3.1 Data Reduction

Two records were removed. One of the records has the name "Travel agency in the park", which is not affiliated to Enron and hence removed from the dataset. Another record with the name "Lockhart Eugene E" was also removed since it has over 17 missing values and thus does not have sufficient information for either training or for being used as a test record.

Five features were removed. The "name" and "email address" features were removed since they are not applicable to other general financial datasets which do not have these specific employees. As mentioned previously, the three features "restricted stock deferred", "loan advances" and "director fees" all have over 85% of their values missing and thus were also removed. Specifically, they have 88%, 98% and 89% of their values missing, respectively.

3.3.2 Splitting into training and test sets

Using stratified splitting, the dataset was split into 70% training set and 30% test set. Both sets have a very similar class distribution, with the training set having roughly 13% of its records as POI and the test set having roughly 12% of its records as POI.

3.3.3 Mean imputation

Due to the abundance of missing values in the dataset, mean imputation was performed. The training set was imputed using mean imputation, which takes the means of each feature and imputes these means into the missing values for the corresponding features. These same means are also imputed into the test set. This is to ensure that the distributional assumptions of the predictive models hold when the model is applied to the test set.

3.3.4 Preprocessed data

After data preprocessing, two sets of data were obtained: 70% training set and 30% test set after stratified splitting with both sets imputed using means from the training set.

4. Fraud Detection Model

4.1 Model Training

Six machine learning algorithms were used to train the model, including decision tree, artificial neural network, support vector machine, naive Bayes, random forest and adaptive boosting.

Decision tree is a tree-based model using decision rules to split records into multiple nodes to classify records. Decision trees however have the tendency to overfit data.

Artificial neural network is an interconnected system of input, hidden and output layers using artificial neurons (units) and weights to process data. Our neural network has 6 units in the hidden layer, a weight decay of 0.01 and a maximum number of iterations of 50.

Support vector machine (SVM) is a kind of generalized linear classifier that classifies binary data according to supervised learning. Its decision boundary is the maximum margin hyperplane for separating distinct classes.

Naive Bayes is a probabilistic model based on Bayes' theorem. It uses conditional probabilities for classification, with the assumption that every feature is conditionally independent of each other.

Random forest is an ensemble learning method which builds multiple decision trees, each with a number of randomly sampled predictors used for decision rules. The majority vote of all trees is used for the final classification. By using an ensemble of classifiers, it reduces overfitting of the final classifier. For our model, we chose the following parameters: 5 (i.e. log(N)+1) variables randomly sampled as candidates each split and 500 trees.

Adaptive boosting is an ensemble learning method with sample weights readjusted at each iteration based on the weight updating coefficient α. Our adaptive boosting algorithm uses 100 maximum iterations, a maximum depth of 3 and "Breiman" weight coefficient calculation ($\alpha = ½*(\ln((1-error)/error))$.

4.2 Standardization

For the support vector machine and neural network methods, the training and test sets are both standardized. Similar to the imputation process, the training set was standardized first and then the same standardization parameters were used to standardize the test set, allowing distribution assumptions of the models to hold on the test set.

Only these two methods require standardization as support vector machines are distance-based, in which features with large scale will dominate other features, and neural networks have accelerated training from standardized inputs.

The other methods used are tree-based or probability-based models, which are both scale invariant and do not require standardization of features.

4.3 Cross Validation

For each type of method, 5-fold cross validation was performed on the training data. The reason for using k-fold cross validation is because it is effective for reducing bias in the presence of small data and can also reduce overfitting by training on multiple training-test splits instead of just one split. The number of folds was chosen to be 5 arbitrarily so that the size of the validation fold is not too small.

During each cross validation iteration, one data balancing technique, either Random Oversampling Example (ROSE) or Synthetic Minority Oversampling Technique (SMOTE), was applied to the training fold. Data

balancing was applied during cross validation instead of before to ensure that there is no data leak, preventing bias in test performance.

The performance metric used to decide the best performing cross validation model is the area under Receiver Operating Characteristic (ROC) curve. The ROC curve is a line graph of true positive rate versus false positive rate at different classification thresholds. The area under ROC ranges from 0 to 1, with 1 being a perfect classifier, 0.5 being a random classifier and 0 being completely wrong. The cross validation model with the highest area under the ROC curve out of the 5 iterations is chosen as the final model.

Since two data balancing techniques were used, each of the 6 machine learning methods has 2 models, one using ROSE and the other using SMOTE. In total, 12 models were made.

4.4 Results

The trained models are used to make predictions on the test set. Recall, which takes true positives and false negatives into account, is the major metric to evaluate the performance of the models because we would like to prioritize detecting POIs. Higher the recall, higher the probability that the model can detect POIs.

	Accuracy	Precision	Recall
Decision Tree	0.881	NA	0
ANN	0.857	0.333	0.2
SVM	0.762	0.222	0.4
Naive Bayes	0.857	0	0
Random Forest	0.857	0	0
Adaboost	0.881	NA	0

The figure on the right shows the prediction results using ROSE. 4 models have recall=0. 2 models have precision=0 and 2 models have precision=NA (i.e. TP+FP=0). It is observed that, although the models show high accuracy, their precision and recall are low. It is caused by 2 reasons. One reason is they have poor predictive power on the positive class. Another reason is the test set is imbalanced (majority is negative). Here, the SVM model achieves the highest recall, 0.4.

	Accuracy	Precision	Recall
Decision Tree	0.714	0.182	0.4
ANN	0.857	0.333	0.2
SVM	0.714	0.111	0.2
Naive Bayes	0.833	0	0
Random Forest	0.857	0.4	0.4
Adaboost	0.857	0.4	0.4

The figure on the second right shows the prediction results using SMOTE. Again, the models show high accuracy but low precision and recall. But this time, there is only 1 model having precision=0 and recall=0. Here, decision tree, random forest and adaboost model achieve the highest recall, 0.4.

In total, there are 4 models achieving the highest recall: SVM using ROSE, decision tree using SMOTE, random forest using SMOTE and adaboost using SMOTE. In order to identify the best model among the 4, tie breakers are needed.

The first tie breaker is precision, which takes true positives and false positives into account. The higher the precision, the lower the chance of false alarms. Given the same recall, it is always good to have a model with lower chance of false alarms. Among the 4 models, there are 2 models achieving the highest precision: random forest using SMOTE and adaboost using SMOTE. In order to identify the best model among the two, one more tie breaker is needed.

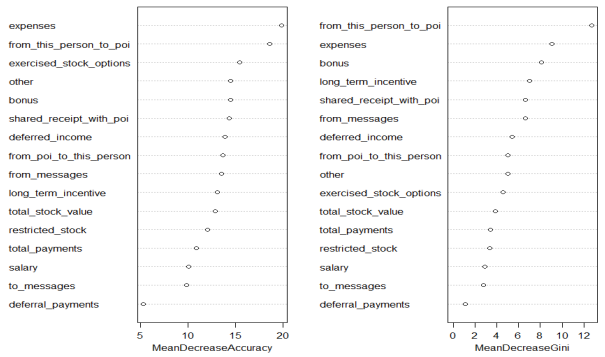
The second tie breaker is training time, which is proportional to the computation cost. Given the same predictive performance (i.e. recall and precision), it is always good to have a model with lower training time to ensure cost-effectiveness. It is observed that the training time for random forest using SMOTE is 1.36 seconds and the training time for adaboost using SMOTE is 20.99 seconds. Therefore, it is concluded that our best model is a random forest using SMOTE.

4.5 Interpretation

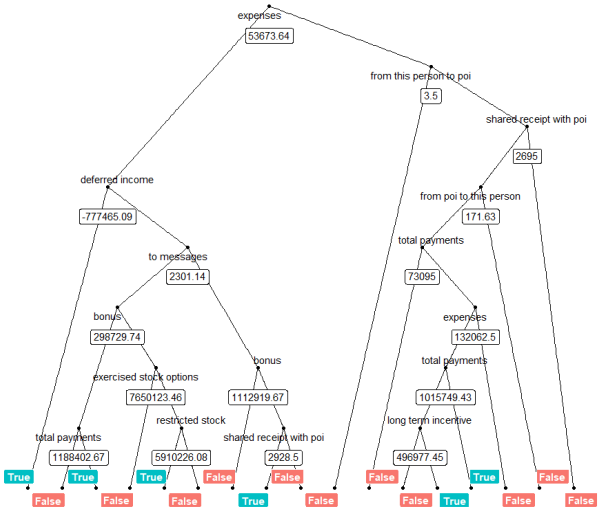
Next, we look into our best model - random forest using SMOTE and interpret its result. The figure on the right shows the confusion matrix of the model applied to the test set. It is observed that the model can predict 2 out of 5 POIs but it also generates 3 false alarms.

	Reference	
Prediction	False	True
False	34	3
True	3	2

The figure on the right shows the variable importance using MDA and MDI. It is observed that, for both MDA and MDI, the top 2 variables are expenses and from_this_person_to_poi while the bottom 3 variables are salary, to_messages and deferral_payments. Therefore, expenses and from_this_person_to_poi have the most predictive power. Also, deferral_payments which achieve comparatively low importance in both MDA and MDI can be discarded.



The figure on the right visualizes the behavior of a decision tree in the random forest. For example, rows with expenses > 53673.64 and from_this_person_to_poi < 3.5 are considered as negatives (i.e. non-POI) and rows with expense < 53673.64 and deferred_income < -777465.09 are considered as positives (i.e. POI).



5. Recommendation

5.1 Non-data analytics element

There are four non-data analytics elements that can be considered. Firstly, lack of corporate control. The main reason why Enron shutdown is mainly due to the immoral behaviors of senior managers, for example, they were falsely reporting the past accounting records and were excepted from the conflicts of interest policy. Thus, they could perform unethical operations that conflict of interest and harm the company without control. Secondly, using unprofessional and aggressive accounting practices to hide the debts and losses. There was an off-balance sheet issue that Enron was using special purpose vehicles (SPVs) to hide its mountains of debt and toxic assets liabilities from investors and creditors. Also, they were using mark-to-market (MTM) techniques that overestimate the future earnings. It misled the investors so they cannot evaluate the potential liabilities of the company properly with the company performance report. Thirdly, bad corporate culture issues as they created toxic corporate culture that led to corruption, greed and deception existed within the

company. Finally, the independence issue of auditors and directors. Enron consulted Arthur Andersen using accounting practice to hide the liabilities of the firm.

5.2 Suggestions

In order to prevent the similar cases happening again, we suggest the following precautions should be taken:

Firstly, the board of directors is required to closely supervise the behavior of senior managers. Corporate governance is critical for a company that refers to a set of processes, practices, policies, laws and systems that affect the way the company operates, manages and controls. At the same time, corporate governance also includes the relationship between many stakeholders and the goal of corporate governance. Enron can implement it in many ways, such as requiring Enron to set up an audit committee to review the work of external auditors and ensure that financial statements are presented fairly and accurately. Or we can ask the board of directors to follow the best application rules of the Cadbury Report. This report provides suggestions on the arrangement of the company's board of directors and accounting system, thereby reducing the risks and shortcomings of corporate governance which have been widely adapted by the different industries over the world.

Secondly, strengthening directors and auditors' independence. Andersen was the world's top five accounting firms providing audit business for Enron, it covered up and concealed Enron's accounting fraud. As such, we should prohibit auditors of listed companies from consulting them at the same time, including maintaining financial data, designing and implementing financial information systems, asset evaluation or valuation services, auditing services, internal audit, etc.

Thirdly, educate, promote and maintain integrity of corporate value. In the Enron case, corporate culture played a critical role, it created toxic corporate culture by using corruption, greed and deception. With this toxic culture, it motivated the executives to use high risk accounting practices to hide the loss of the company so as to get high investment return. It is understandable to motivate the employees to help the company to earn more revenue, but at the same time, it is necessary to make sure the code of conduct and ethics. Several things we can do, for example, conduct compulsory training regularly to educate and promote integrity culture, creating corporate rules to penalize the employees who conduct misconducted behaviors.

6. Summary

Aiming to identify the POIs from Enron, we have built different models using various algorithms and 2 different imbalanced dataset handling techniques. After the study, it is found that the best model is built using random forest.With SMOTE balancing method, it can achieve 86% accuracy, 40% precision and 40% recall on the testing dataset. With this fraud detection model, we can use it to identify persons of interest with the extracted features from new email history records or voice recording in the organization without manual involvement. In the future planning, we can also perform social network analysis on the identified potential person of interest to find related parties for further investigation. Apart from the quantitative way for fraud detection, we suggested strengthening directors and auditors independently, promoting and ensuring integrity of corporate value, and requiring the board of directors to closely supervise the behavior of senior managers for fraud prevention.

Reference

1. *Plotting trees from random forest models with ggraph*. bbsmax. (n.d.). Retrieved December 3, 2021, from https://www.bbsmax.com/A/kPzOXYjoJx/.
2. Segal, T. (2021, December 1). *Enron scandal: The fall of a wall street darling*. Investopedia. Retrieved December 3, 2021, from https://www.investopedia.com/updates/enron-scandal-summary/.
3. Swalin, A. (2018, March 19). *How to handle missing data*. Medium. Retrieved December 3, 2021, from https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4.
4. *Data science 101: Normalization, standardization, and regularization*. KDnuggets. (n.d.). Retrieved December 3, 2021, from https://www.kdnuggets.com/2021/04/data-science-101-normalization-standardization-regularization.html.
5. missmariss31. (n.d.). *Missmariss31/Enron: Udacity Machine Learning Project using the enron financial and email dataset*. GitHub. Retrieved December 3, 2021, from https://github.com/missmariss31/enron.
6. *The cadbury report*. Cambridge Judge Business School : The Cadbury Archive : The Cadbury Report. (n.d.). Retrieved December 3, 2021, from http://cadbury.cjbs.archios.info/report.
7. Desarda, A. (2019, January 17). *Understanding adaboost*. Medium. Retrieved December 3, 2021, from https://towardsdatascience.com/understanding-adaboost-2f94f22d5bfe.
8. Dharan, B. G., & Bufkins, W. R. (2008, July 24). *Red Flags in Enron's reporting of Revenues & Key Financial Measures*. SSRN. Retrieved December 3, 2021, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1172222.
9. *Innovation corrupted: How managers can avoid another Enron*. HBS Working Knowledge. (2008, July 7). Retrieved December 3, 2021, from https://hbswk.hbs.edu/item/innovation-corrupted-how-managers-can-avoid-another-enron.

Appendix

Figure 1. Histograms of each numerical variable