MINI CASE STUDY: REAL-LIFE FRAUD DETECTION SCENARIO

Shihao Liang
[1]The University of Hong Kong [2]FITE-7410 Mini-case Study
shihaol@connect.hku.hk

ABSTRACT

Objective: The primary objective is to develop a machine learning model capable of detecting fraudulent transactions in the ENRON dataset. This entails identifying patterns and anomalies that differentiate fraudulent activities from normal transactions. **Scope**: There are four scopes, a) data analysis: Analyze the dataset to understand the nature of transactions, including their frequency, amounts, and other relevant features, b) feature identification: Determine which features or combinations of features are most indicative of fraud, c) model development: Develop a predictive model using machine learning techniques to identify potential fraudulent transactions, and d) validation: Validate the model's effectiveness using appropriate metrics and methodologies.

1 EXPLORATORY DATA ANALYSIS

The ENRON dataset contains various financial and email communication features for different individuals. Here's an initial overview of its structure:

- It includes columns such as *salary*, *to_messages*, *deferral_payments*, *total_payments*, *bonus*, *email_address*, *restricted_stock_deferred*, and many more.
- The *poi* column indicates whether the individual is a person of interest in the fraud case (*True* or *False*).
- The dataset has a mix of numerical and categorical data, and there are missing values in several columns.

The Exploratory Data Analysis (EDA) includes four steps: Statistical Summary, Data Visualization, Identifying Predictors for Fraud and Data Cleaning and Preprocessing. We start with the statistical summary and then proceed to data visualization and further analysis. We also identify any immediate data cleaning needs as we dive in.

1.1 STATISTICAL SUMMARY

The statistical summary provides a comprehensive view of the dataset. Here's a brief overview:

Numerical Features:

- *Salary:* Varies widely, with a maximum of around $26.7 million.
- *To Messages:* Average of about 2074 messages, but with a large standard deviation.
- *Deferral Payments* and *Total Payments:* Show significant variation, indicating diverse financial arrangements.
- *Bonus:* Also varies considerably, with some extremely high bonuses.
- *Exercised Stock Options:* Some individuals have very high values, possibly a key feature to look at for fraud detection.
- *From Messages:* Shows wide variation in email activity.

Categorical Features:

- *Email Address:* Unique for each individual.

Observations:

- The dataset contains a mix of financial and email communication features.
- There are significant outliers in many financial features, which could be indicative of fraudulent activities.
- The presence of missing values in several columns will require careful handling during data preprocessing.

Table 1 shows the key numerical features, which could be utilized for following experiments and case study.

Feature	**Mean**	**Median**	**Standard Deviation**
Salary	562194.29	259996.00	2716369.15
To Messages	2073.86	1211.00	2582.70
Deferral Payments	1642674.15	227449.00	5161929.97
Total Payments	5081526.49	1101393.00	29061716.40
Bonus	2374234.61	769375.00	10713327.97

Table 1: Statistical Summary of Key Numerical Features

1.2 Data Visualization

The visualization methods chosen are histograms and scatter plots.

Histograms for distribution of key numerical features. Results are shown in 1. **Salary, Bonus, Total Payments, and Exercised Stock Options**, these features have a right-skewed distribution, indicating that a few individuals have significantly higher values compared to the rest. Such distributions are common in financial datasets but also can be indicative of outliers or unusual transactions. **From Messages and To Messages** are also skewed, suggesting varied levels of email activity among individuals.

Scatter plots or correlation matrices to understand relationships between variables. Results are shown in 2. The scatter plots with the poi (Person of Interest) variable show how these financial and email features interact with the classification of individuals as $POIs$.

There appear to be differences in the distribution of these features when comparing $POIs$ to non $POIs$, particularly in features like **Bonus, Total Payments, and Exercised Stock Options**.

1.3 Identifying Predictors for Fraud

Correlation analysis. Correlation matrix is shown in Figure 3. The correlation of features with the POI variable shows: $shared\_receipt\_with\_poi$, $from\_poi\_to\_this\_person$, and $from\_this\_person\_to\_poi$ have the most significant positive correlations with POI. Features like $deferral\_payments$ show a negative correlation. It's important to note that correlation does not imply causation, but these relationships can provide insights into which features are most associated with the POI variable.

1.4 Data Cleaning and Preprocessing

This is the data preparation process of the following model training and evaluation. For data cleaning and preprocessing, we:

- **Handling Missing Values:**
  - Set a threshold for missing values (80%).

  - Calculate and display the percentage of missing values per column.
  - Drop columns with a missing value percentage above the threshold.
- **Dropping Irrelevant Features:** Remove the 'email_address' column as it is not useful for modeling.
- **Segmentation by POI:** Split the dataset into two subsets based on the POI.
- **Data Sampling and Augmentation:**
  - Define size parameters for training and test datasets.
  - Randomly sample a subset of 'poi_false_data' and 'poi_true_data' for the test set.
- **Preparing Training and Test Datasets:**
  - Use the rest of the data as train set.
  - Convert the response variable to a factor.

2 Model Building and Validation

2.1 Model Selection

Given the nature of the dataset (classification problem), we start with these two algorithms:

- **Random Forest Classifier:** This is an ensemble learning method that is good for handling a mix of numerical and categorical data, and is robust to outliers.
- **Logistic Regression:** A fundamental classification algorithm that's useful for binary classification problems like fraud detection.

2.2 Data Splitting

We first follow the preprocessing procedure as stated in Section 1, and then split the dataset into training and testing sets, with a common split ratio 80% for training and 20% for testing.

2.3 Model Evaluation

We evaluate the models using metrics like:

- **Accuracy:** Measures the overall correctness of the model.
- **Precision and Recall:** Particularly important in fraud detection to balance the trade-off between catching as many fraud cases as possible (recall) and ensuring that the predictions are accurate (precision).
- **F1-Score:** Harmonic mean of precision and recall.
- **ROC-AUC Score:** Measures the performance across all possible classification thresholds.

2.4 Model Performance and Analysis

The performance metrics for both the Random Forest and Logistic Regression models on the test set are presented in Table 3.

Algorithm comparison. The Random Forest model shows high accuracy and precision but has a moderate recall. This means it is very accurate in the predictions it makes, but it may miss half of the actual fraud cases.

The Logistic Regression model, on the other hand, has decent accuracy but fails to identify any true positives (as indicated by the recall and precision of 0%). This could be due to the model not being complex enough to capture the nuances of the dataset or due to class imbalance.

Iterate process. We iterate on the random forest algorithm with hyper-parameter grid search. Table 2 shows the search range, and the best parameter combination is: $n.trees = 100$, $interaction.depth = 10$, $n.minobsinnode = 5$. The iterate process helps improving performance from 50.40%(lowest) to 66.67%(highest).

n.trees	**interaction.depth**	**n.minobsinnode**
50	**10**	2
100	20	**5**
150	30	10

Table 2: Hyper-parameter grid search ranges. The best parameters is bold.

Feature importance. The detailed feature importance results are shown in Table 5 in the appendix.

The top features, shown in 4, based on importance from the Random Forest model are: $deferred\_income$, $other$, $exercised\_stock\_options$, $total\_stock\_value$ and $bonus$. These features have the highest importance scores, suggesting they are most influential in predicting whether an individual is a POI in the fraud case.

Metric	Random Forest	Logistic Regression
Accuracy	96.67%	83.33%
Precision	100.00%	0.00%
Recall	50.00%	0.00%
F1 Score	66.67%	0.00%
ROC AUC	82.14%	44.64%

Table 3: Best performance using random forest and logistic regression.

Feature	**Importance**
deferred_income	0.0905
other	0.0826
exercised_stock_options	0.0816
total_stock_value	0.0761
bonus	0.0741

Table 4: Feature Importances from the Random Forest Model, sorted in descending order of importance.

3 Fraud Scenario Identification

3.1 Scenario Description

Consider a scenario where an individual, referred to as Employee X, is suspected of financial fraud within the company. The suspicions arise due to the following observations:

- **Unusually High Deferred Income:** Employee X has deferred income significantly higher than the average for their position.
- **Large Amount of Stock Options Exercised:** Employee X recently exercised a large number of stock options, which is unusual given their role and tenure in the company.
- **High Bonus Relative to Salary:** The bonus received by Employee X is disproportionately high compared to their salary.
- **Frequent Communication with Known POIs:** Employee X has been frequently communicating with known Persons of Interest (POIs) in the ongoing fraud investigation.

3.2 Data-Informed Decision Making

Using the trained Random Forest model, we can input these observations to make a data-informed decision. The model, which prioritizes features like $deferred_income$, $exercised_stock_options$, and

$bonus$, can help assess the likelihood that Employee X is involved in fraudulent activities. So, we create a hypothetical feature set for Employee X based on the scenario and use the model to predict the likelihood of them being a POI. Detailed steps and results are shown below.

Feature Extraction. First, we extract the relevant features for Employee X from the fake data. The extracted features and their values are shown in Table 6.

Model Prediction. Use the trained Random Forest model to predict whether Employee X is likely to be a POI based on these features. The Random Forest model classifies Employee X as **Low Risk** for being a POI in the context of financial fraud.

Analysis. Despite crafting a data point with attributes that are typically considered red flags for fraud, the model's assessment is Low Risk. This outcome could be due to several reasons:

- Data and Feature Limitations: The model's training data might not contain enough examples of similar fraudulent patterns, limiting its ability to recognize such cases.
- Complex Interactions: The model considers interactions between various features. The combination of features in X profile, although individually suspicious, might not align with the fraud patterns learned by the model.
- Threshold for Detection: The model might have a high threshold for classifying an individual as high risk, possibly to minimize false positives.

Reflections on the Model's Prediction First, the model sensitivity. This result suggests revisiting the model's sensitivity and possibly recalibrating it to be more attuned to subtle signs of fraud. Then, a feature re-evaluation is important. There might be a need to include more predictive features or to re-evaluate the feature engineering process. Last but not least, we need a continued vigilance. Despite the model's prediction, in a real-world scenario, X profile would still warrant further investigation due to the combination of red flags.

3.3 Caveats and Further Actions

Model Limitations. The model's prediction is as good as the data and features it has been trained on. It's possible that the model may not capture all nuances of fraud or may have biases based on the training data.

Manual Review. Given the high stakes of fraud detection, a manual review of Employee X's activities, especially considering the unusual patterns noted, would be prudent. The model's prediction should be used as a guide, not a definitive judgment.

Continuous Monitoring. If Employee X's financial activities continue to exhibit unusual patterns, continuous monitoring and periodic reassessment using updated data and models would be advisable.

4 Non-data analytic element:

4.1 Risks and Red Flags

Unusual Financial Transactions: Large, irregular transactions or changes in financial patterns that do not align with business activities.

Conflicts of Interest: Situations where employees' personal interests could conflict with those of the organization.

Lack of Transparency: Inadequate disclosure of financial information and decision-making processes.

Weak Internal Controls: Ineffective or insufficient internal controls and audit processes.

Management Override: Situations where management can easily override controls for personal gain or to manipulate financial results.

4.2 Corporate Governance and Controls

Strong Ethical Culture: Establishing a strong culture of ethics and integrity at all levels, emphasizing the importance of ethical behavior.

Effective Oversight Structures: Ensuring that oversight bodies like the Board of Directors are independent, well-informed, and actively oversee management and financial reporting.

Robust Internal Controls: Implementing and regularly reviewing internal controls over financial reporting and operations.

Transparent Reporting: Ensuring transparency in financial reporting and corporate disclosures.

Whistleblower Policies: Encouraging and protecting whistleblowers who report suspicious activities.

4.3 Suggestions to Prevent Similar Frauds

Regular Audits and Reviews: Conduct regular external and internal audits to review financial and operational processes.

Training and Awareness Programs: Regular training for employees on ethics, fraud awareness, and the importance of internal controls.

Enhanced Monitoring Mechanisms: Implement advanced monitoring systems to detect unusual transactions and behaviors indicative of fraud.

Conflict of Interest Policies: Establish clear policies to manage and disclose conflicts of interest.

Risk Management Framework: Develop a comprehensive risk management framework that identifies, assesses, and manages financial and operational risks.

5 Conclusion

Preventing financial fraud requires a combination of strong corporate governance, effective internal controls, an ethical work culture, and active engagement from all stakeholders, including employees, management, and the board. Data analytics is a powerful tool in this endeavor, but it should be complemented by these broader organizational and cultural measures.

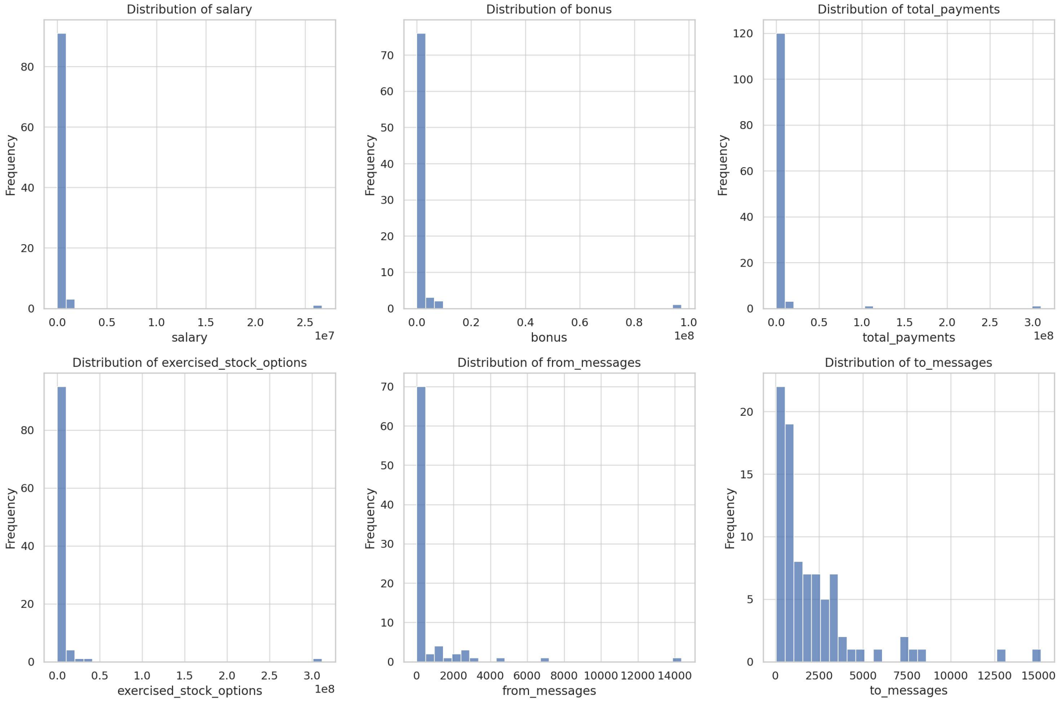

Figure 1: Histograms for key numerical features.

Appendix

A Implementation Details

Feature	Importance
deferred_income	0.0905
other	0.0826
exercised_stock_options	0.0816
total_stock_value	0.0761
bonus	0.0741
restricted_stock	0.0730
shared_receipt_with_poi	0.0695
total_payments	0.0637
expenses	0.0609
from_this_person_to_poi	0.0556
long_term_incentive	0.0550
salary	0.0481
from_poi_to_this_person	0.0430
to_messages	0.0423
from_messages	0.0401
deferral_payments	0.0317
restricted_stock_deferred	0.0087
loan_advances	0.0028
director_fees	0.0006

Table 5: Feature Importances from the Random Forest Model, sorted in descending order of importance.

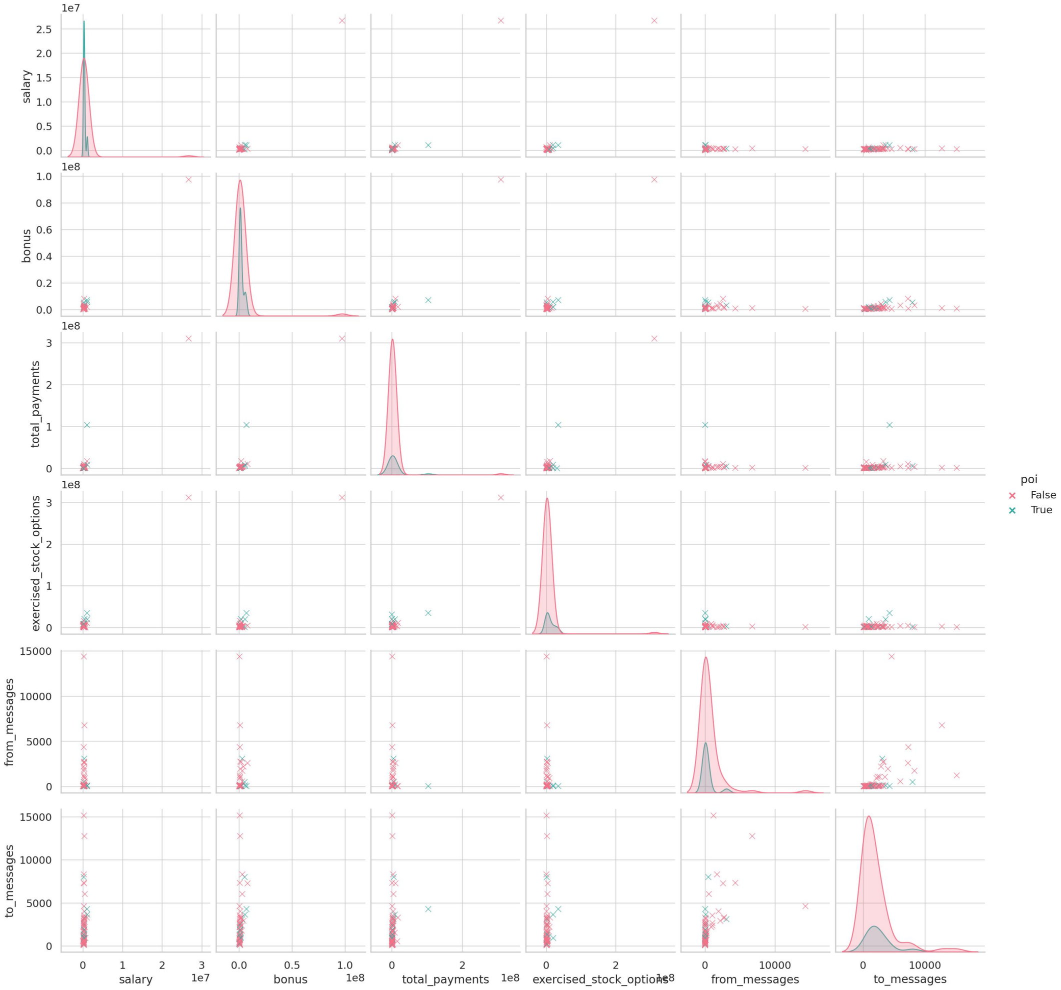

Figure 2: Scatter plots to examine relationships with the poi variable.

References

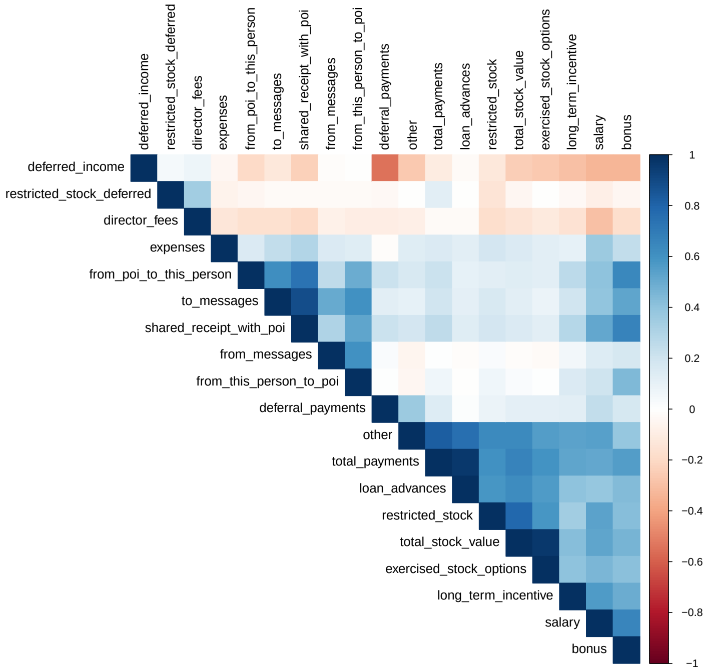

Figure 3: Scatter plots to examine relationships with the poi variable.

Feature	Value
salary	300000 (Higher than average salary)
to_messages	1000 (Average number of to-messages)
deferral_payments	2000000 (Unusually high deferred income)
total_payments	800000 (Total payments including salary, bonus, etc.)
loan_advances	0 (Assuming no loan advances)
bonus	1500000 (High bonus relative to salary)
restricted_stock_deferred	0 (Assuming standard restricted stock deferred)
deferred_income	-1000000 (High negative deferred income, indicating a potential red flag)
total_stock_value	2000000 (High stock value)
expenses	50000 (General expenses)
from_poi_to_this_person	100 (High frequency of communication with POIs)
exercised_stock_options	1500000 (Large amount of exercised stock options)
from_messages	200 (Average number of from-messages)
other	100000 (Other financial benefits)
from_this_person_to_poi	50 (Frequent communication to POIs)
long_term_incentive	250000 (Long-term incentives)
shared_receipt_with_poi	500 (Shared receipt with POI)
restricted_stock	500000 (Restricted stock value)
director_fees	0 (Assuming no director fees)

Table 6: Employee X Feature Set.