UTILIZING MACHINE LEARNING TO IDENTIFY FRAUDULENT ACTIVITIES THROUGH ANALYSIS OF FINANCIAL AND COMMUNICATION DATA

Xiaojian Lin[1], Zhaochen Li[2], Xiwen Liang[3], Jialin Liu[4], Shaochong Yan[5]
[1]ID:3036196544
[2]ID:3036195071
[3]ID:3036196544
[4]ID:3036196544
[5]ID:3036196544

Department of Computer Science, The University of Hong Kong

ABSTRACT

Financial fraud, necessitating sophisticated detection mechanisms significantly undermines the stability of economic systems. This study presents a comprehensive analysis using the Enron dataset to pinpoint predictors of fraudulent activities. Employing exploratory data analysis (EDA), we dissect the dataset's complex structure and statistical nuances, applying data visualization and statistical methodologies to reveal the interplay between various variables and occurrences of fraud. Intensive data cleansing and preprocessing underpin this exploration.

Central to the research is the evaluation of eight machine learning (ML) models: Support Vector Machines (SVM), AdaBoost, Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbors (KNN), Gradient Boosting Machines (GBM), Decision Trees (DT), and Neural Networks (NN). Gradient Boosting Machines (GBM) and Random Forest (RF), recognized for their predictive precision, outperform others in identifying fraudulent transactions. We also compare our results with other open source implementations, and some improvements have been made on the existing research to make our model have higher prediction accuracy.

Beyond algorithmic strategies, the investigation extends to real-world fraud scenarios, capitalizing on the predictive prowess of advanced models for informed decision-making. It also highlights the importance of integrating non-analytical factors, such as corporate governance, into fraud risk mitigation strategies.

***Index Terms*—**Financial Fraud Detection, Enron Financial Dataset, Fraudulent Activities Predictors, Exploratory Data Analysis (EDA), Data Visualization Techniques, Statistical Analysis in Fraud Detection, Data Cleansing and Preprocessing, Machine Learning Models Evaluation, Gradient Boosting Machines (GBM), Random Forest (RF)

1. INTRODUCTION

1.1 Background

The specter of financial fraud has long cast a shadow over the integrity of economic systems worldwide. The infamous Enron scandal serves as a stark reminder of the cataclysmic consequences that can ensue from such deceptions. In the aftermath, the impetus to develop sophisticated detection mechanisms has intensified, particularly within the domain of financial analytics. The advent of machine learning (ML) offers a promising frontier in this endeavor. By leveraging computational techniques capable of identifying subtle patterns indicative of fraudulent activity, ML provides a crucial tool in the arsenal against financial malfeasance. In this digital era, ML transcends traditional analytics, offering an advanced shield against the intricate schemes of financial fraudsters.

Objectives

The primary objective of this case study is to harness the advanced capabilities of machine learning (ML) for detecting fraudulent transactions within the real-world dataset from the Enron case. We aim to conduct a comprehensive exploration of the dataset, focusing on an in-depth analysis of its structure, features, and statistical properties. Through meticulous exploratory data analysis (EDA), we aim to uncover the intricate relationships between various variables and fraudulent outcomes, identifying pivotal predictive factors and risk indicators. This process of distillation of complex data into clear indicators of fraudulent behavior enables us to develop more targeted and effective detection strategies.

In pursuit of this goal, we have employed a diverse set of eight ML models: Neural Networks, AdaBoost, Random Forest, Support Vector Machines (SVM), Logistic Regression, K-Nearest Neighbors, Gradient Boosting Machines, and Decision Trees. Each model was meticulously calibrated and optimized through rigorous parameter tuning. Despite the variety of approaches,

Gradient Boosting Machines (GBM) and Random Forest (RF) stood out for their exceptional performance in detecting fraudulent activities. These two models were therefore selected for an in-depth exploration, considering their superior efficacy.

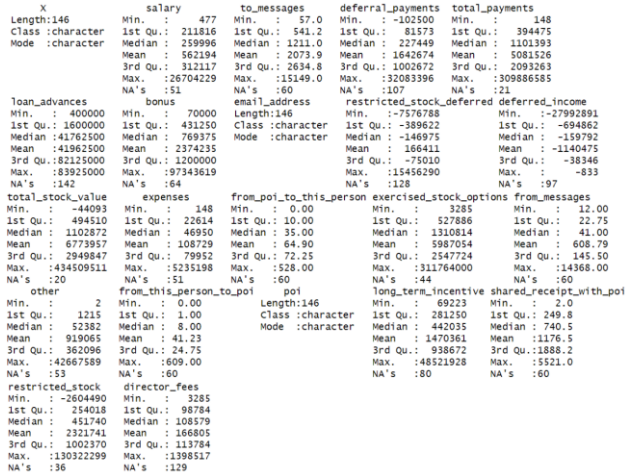
Gradient Boosting Machines (GBM) are powerful machine learning algorithms known for their effectiveness in handling complex datasets by sequentially building and optimizing decision trees to improve predictive accuracy [1, 2, 3]. And Random Forest model is known for its ensemble approach that combines multiple decision trees to enhance decision accuracy and prevent overfitting [4, 5, 6]. The combination of these two models with the best performance lays a good foundation for our fraud detection work in financial scenarios.

In this study, we apply advanced machine learning techniques, particularly Gradient Boosting Machines (GBM) and Random Forest (RF), to detect fraud in the Enron dataset. Our focus is on comprehensive data analysis and model optimization to identify key indicators of fraudulent behavior. GBM and RF stand out for their accuracy in detecting fraud, combining complex data handling and decision-making robustness. We also consider non-analytical factors like corporate governance, aiming for a holistic fraud prevention approach. This research provides actionable insights for enhancing financial security and mitigating fraud risks.

2. DATASET DESCRIPTION AND FRAUD ANALYSIS METHODS

2.1 Dataset Overview

The ENRON dataset under examination provides a unique window into the financial records of the eponymous corporation. Also it provides valuable insights into its operations. Although the dataset comprises an incomplete annual report, it encompasses crucial components such as the income statement, balance sheet, and cash flow statement. These financial indicators shed light on the company's performance, highlighting aspects like income, payments, bonuses, email addresses, and stock data. Analyzing this dataset enables us to delve deeper into the intricacies of ENRON's financial landscape and gain a comprehensive understanding of its operations. The content of this ENRON Dataset is mainly an incomplete company annual report. This annual report contains part of the income statement, balance sheet, and cash flow statement. The main data of the dataset includes income, payment, bonus, email address, stock data, etc. Table 1 below is a summary of the data set.



```
       X                salary          to_messages     deferral_payments  total_payments
 Length:146         Min.   :     477   Min.   :   57.0   Min.   : -102500   Min.   :     148
 Class :character   1st Qu.:  211816   1st Qu.:  541.2   1st Qu.:   81573   1st Qu.:  394475
 Mode  :character   Median :  259996   Median : 1211.0   Median :  227449   Median : 1101393
                    Mean   :  562194   Mean   : 2073.9   Mean   : 1642674   Mean   : 5081526
                    3rd Qu.:  312117   3rd Qu.: 2634.8   3rd Qu.: 1002672   3rd Qu.: 2093263
                    Max.   :26704229   Max.   :15149.0   Max.   :32083396   Max.   :309886585
                    NA's   :51         NA's   :60        NA's   :107        NA's   :21
 loan_advances        bonus          email_address       restricted_stock_deferred deferred_income
 Min.   :  400000   Min.   :   70000   Length:146         Min.   :-7576788          Min.   :-27992891
 1st Qu.: 1600000   1st Qu.:  431250   Class :character   1st Qu.: -389622          1st Qu.:  -694862
 Median :41762500   Median :  769375   Mode  :character   Median : -146975          Median :  -159792
 Mean   :41962500   Mean   : 2374235                      Mean   :  166411          Mean   : -1140475
 3rd Qu.:82125000   3rd Qu.: 1200000                      3rd Qu.:  -75010          3rd Qu.:   -38346
 Max.   :83925000   Max.   :97343619                      Max.   :15456290          Max.   :     -833
 NA's   :142        NA's   :64                            NA's   :128               NA's   :97
 total_stock_value    expenses      from_poi_to_this_person exercised_stock_options from_messages
 Min.   :  -44093   Min.   :   148   Min.   :  0.00          Min.   :     3285      Min.   :  12.00
 1st Qu.:   494510  1st Qu.:  22614   1st Qu.: 10.00          1st Qu.:   527886      1st Qu.:  22.75
 Median :  1102872  Median :  46950   Median : 35.00          Median :  1310814      Median :  41.00
 Mean   :  6773957  Mean   : 108729   Mean   : 64.90          Mean   :  5987054      Mean   :  608.79
 3rd Qu.:  2949847  3rd Qu.:  79952   3rd Qu.: 72.25          3rd Qu.:  2547724      3rd Qu.:  145.50
 Max.   :434509511  Max.   :5235198   Max.   :528.00          Max.   :311764000      Max.   :14368.00
 NA's   :20         NA's   :51        NA's   :60              NA's   :44             NA's   :60
    other          from_this_person_to_poi     poi          long_term_incentive shared_receipt_with_poi
 Min.   :      2   Min.   :  0.00          Length:146         Min.   :   69223    Min.   :   2.0
 1st Qu.:   1215   1st Qu.:  1.00          Class :character   1st Qu.:  281250    1st Qu.: 249.8
 Median :  52382   Median :  8.00          Mode  :character   Median :  442035    Median : 740.5
 Mean   : 919065   Mean   : 41.23                             Mean   : 1470361    Mean   :1176.5
 3rd Qu.: 362096   3rd Qu.: 24.75                             3rd Qu.:  938672    3rd Qu.:1888.2
 Max.   :42667589  Max.   :609.00                             Max.   :48521928    Max.   :5521.0
 NA's   :53        NA's   :60                                 NA's   :80          NA's   :60
 restricted_stock    director_fees
 Min.   : -2604490   Min.   :   3285
 1st Qu.:   254018   1st Qu.:  98784
 Median :   451740   Median : 108579
 Mean   :  2321741   Mean   : 166805
 3rd Qu.:  1002370   3rd Qu.: 113784
 Max.   :130322299   Max.   :1398517
 NA's   :36          NA's   :129
```

Table 1: Summary of the ERON Dataset

2.2 Data Cleaning and Preprocessing

The final preparatory phase involves tailoring the data for both EDA and machine learning modeling:

Our initial scrutiny revealed variables with a significant portion of missing values, such as 'loan_advances' and 'director_fees', which were predominantly 'NA's'. Considering their sparse nature, these were candidates for removal to streamline the dataset. For essential variables with missing data, like 'salary' and 'total_payments', we applied imputation techniques to fill in gaps, using methods such as median substitution or model-based approaches, preserving the underlying data distribution.

Outlier detection was another critical step; for example, the 'bonus' variable showed potential outliers with values significantly deviating from the mean. We assessed these outliers in the context of potential fraud indicators versus statistical anomalies, applying robust statistical methods or domain-informed thresholds to either adjust or remove these values.

Through these meticulous cleaning and preprocessing steps, we ensured the dataset's quality and integrity, setting a firm foundation for the subsequent EDA and ML modeling phases, which aim to uncover patterns and predictors of fraudulent behavior within the Enron corpus. The above data cleaning is just some examples. Specifically, in the process of EDA and machine learning model building and analysis, there will certainly be more specific problems related to data, whether it is missing value or abnormal value, which need to be dealt with according to the situation.

2.3 Exploratory Data Analysis (EDA)

Univariate Analysis: We start with a univariate examination of each variable to understand its distribution and individual characteristics. Tools like histograms and boxplots aid in this initial analysis. For instance, Figure 1 shows a histogram of bonuses, providing insights into their distribution across the dataset. Figure 2, a boxplot of

'to_messages', helps identify potential outliers or data inconsistencies.

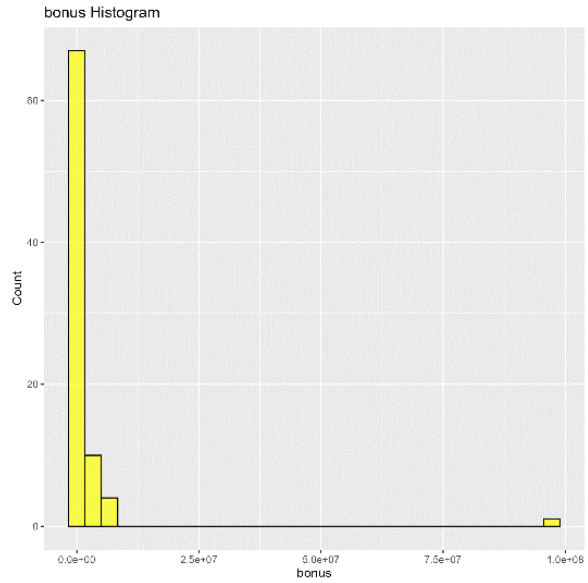


Figure 1: bonus Histogram

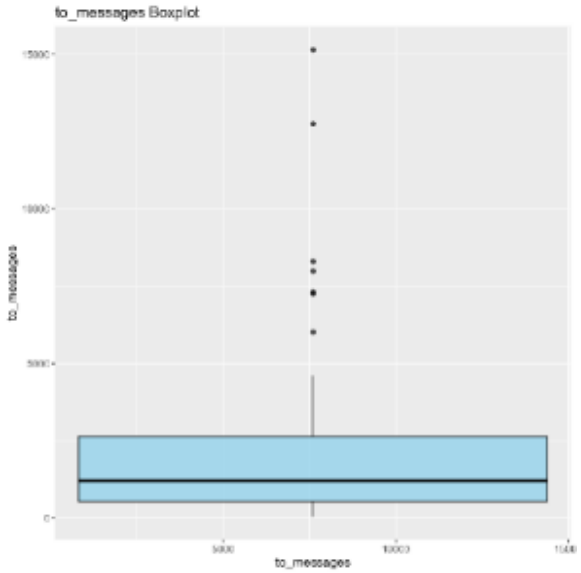
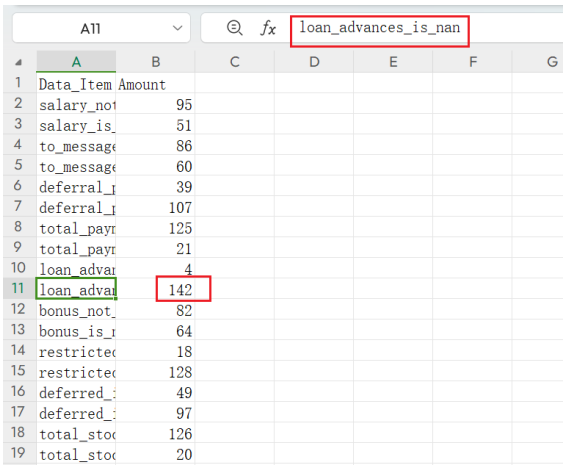


Figure 2: to_message Boxplot

Missing Data Analysis: Integral to our univariate approach is the assessment of missing values within each variable. We evaluate the extent of missing data, as missing entries could significantly impact the quality of our analysis and modeling. For variables with a high incidence of missing values, such as the high value of "loan_advances_is_nan" shown in Figure 3, the variable "loan_advances" has a high degree of missing, and depending on the correlation of the variable and the nature of the missing, it is decided to enter or discard.



	A	B	C	D	E	F	G
1	Data_Item	Amount					
2	salary_not	95					
3	salary_is_	51					
4	to_message	86					
5	to_message	60					
6	deferral_p	39					
7	deferral_p	107					
8	total_payr	125					
9	total_payr	21					
10	loan_advar	4					
11	loan_advar	142					
12	bonus_not_	82					
13	bonus_is_r	64					
14	restricted	18					
15	restricted	128					
16	deferred_i	49					
17	deferred_i	97					
18	total_stoc	126					
19	total_stoc	20					

Figure 3: Part of the Valid and Missing Data Table

Outlier Analysis: Concurrently, we conduct outlier detection to spot anomalous entries that may distort our statistical conclusions. By analyzing boxplots and employing statistical tests, we discern outliers that could indicate data entry errors or genuine but rare events that warrant further investigation. For example, Figure 2's boxplot of 'to_messages' is scrutinized for values that deviate from the norm, which might be indicative of data entry errors or potential signs of fraudulent activity.

Bi-/Multi-variate Analysis: Building on univariate insights, we explore relationships between variables. Scatter plots and correlation matrices are key tools here. Figure 4's scatter plot could, for instance, reveal correlations between two financial metrics, while Figure 5's correlation matrix offers a broader view of interdependencies across multiple variables. This analysis helps in understanding how variables collectively relate to fraudulent activity and in constructing a predictive model with greater accuracy.

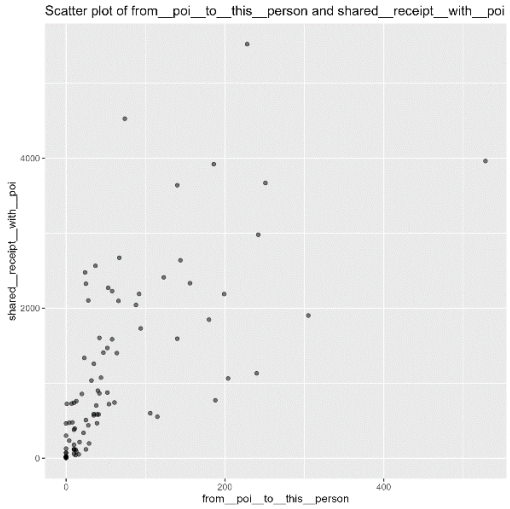


Figure 4: Sample Scatter Plot

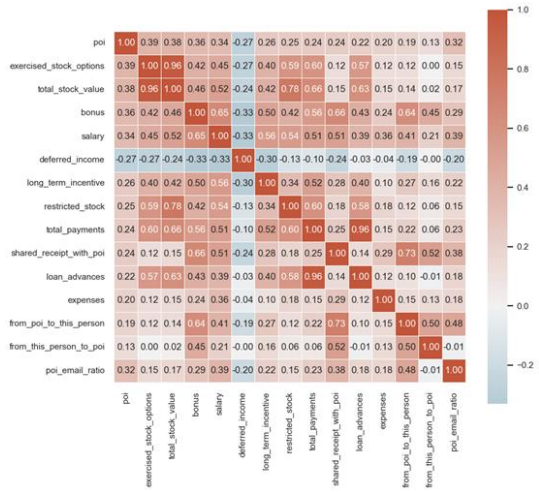


Figure 5: Heatmap of Correlation Coefficients of all Feature Terms

2.4 Identification of Predictive Variables

In this pivotal phase of our study, we aim to pinpoint variables within the Enron dataset that strongly indicate fraudulent activities. This involves a detailed analysis of data patterns and anomalies, focusing on two key aspects:

Key Predictive Factors: Rigorous statistical analysis was employed to identify significant predictors of fraud within the dataset. Our univariate analysis, as demonstrated in Figure 1 (bonus Histogram) and Figure 2 (to_message Boxplot), provided initial insights into the distribution and characteristics of individual financial metrics. This analysis was instrumental in discerning a spectrum of financial figures – from revenues and expenses to other salient metrics – and served as the foundation for identifying pivotal financial indicators. These key predictive factors are crucial in developing a nuanced understanding of the fraud risks associated with various aspects of the dataset.

Risk Signal Detection: The second focus of our analysis was on detecting risk signals or red flags within the data that could indicate potential fraudulent activities. We scrutinized data patterns and anomalies that deviated from expected norms, applying both visual and statistical methods. The histograms and boxplots used in our univariate analysis not only highlighted the typical data distributions but also revealed outliers and unusual patterns that might signal fraudulent activities. By examining these visual representations, we could identify unusual spikes or dips in financial figures, such as unusually high bonuses or irregular messaging patterns, which are often indicative of underlying fraud.

The combination of these two analytical approaches – identifying key predictive factors and detecting risk signals – provided a comprehensive method for uncovering and understanding the indicators of fraudulent behavior in the Enron dataset. This holistic approach is crucial for developing effective fraud detection models that can be applied in real-world scenarios to prevent similar financial malfeasance.

3. MODEL BUILDING AND VALIDATION

3.1 Model Selection

3.1.1 Introduction to Model Selection

In addressing the complexities of financial fraud detection, we strategically selected multiple machine-learning models for our study. The Enron dataset, typical of financial data, presents high-dimensional features and intricate patterns, necessitating an approach that goes beyond single-model analysis. Employing a variety of models allows us to explore the dataset's multifaceted nature thoroughly and effectively detect fraud indicators.

3.1.2 Initial attempt on decision tree

We initially tried to use decision trees to identify fraud, and the decision tree model can effectively deal with nonlinear relationships and complex feature interactions. In financial fraud analysis, fraudulent behavior often exhibits non-linear feature associations, and decision trees can capture these complex patterns. Furthermore, the Enron data set mentioned in the previous section may have a large number of features and high dimensions, including both categorical variables, such as fraud and non-fraud labels, and numerical variables, such as transaction amounts. The decision tree model can not only handle such high-dimensional data without feature selection or dimensionality reduction operations, but also can handle both types of variables and automatically select the best segmentation point.

3.1.3 Follow-up exploration on GBM and RF

While decision trees have some advantages for analyzing the Enron dataset, they also have certain limitations. For example, decision trees are prone to overfitting, especially when dealing with complex and high-dimensional datasets. They may focus too much on noise and local features, resulting in poor generalization performance on new data. Due to their tendency to overfit, decision trees may perform well on the training data but poorly on new data. This indicates that decision trees may have high variance and be overly sensitive to noise and randomness in the data. It is also noteworthy that decision trees tend to favor predicting the majority class and may overlook predictions for the minority class. This makes it perform poorly on some unbalanced datasets, especially Enron datasets.

In the follow-up exploration, we found that Gradient Boosting Machines (GBM)and Random Forest(RF) can well solve the problems of overfitting and data imbalance in decision tree model. Compared with decision tree, GBM excels in improving predictive accuracy through

sequential model building and optimization, making it highly effective in identifying. subtle fraud patterns. Random Forest, an ensemble method, is known for its robustness in handling high dimensional data and providing valuable insights into feature importance. This ensemble approach helps in preventing overfitting while maintaining high accuracy, crucial for reliable fraud detection.Therefore, GBM and RF are our models of choice.

3.1.4 Overview of Other Models

The remaining models, including AdaBoost, SVM, Neural Networks, Logistic Regression, and K-Nearest Neighbors, complement our primary choices. AdaBoost adapts to complex patterns by concentrating on challenging classifications, while SVM effectively handles high-dimensional spaces. Neural Networks are capable of revealing intricate non-linear relationships, and Logistic Regression offers clear interpretability. K-Nearest Neighbors is excellent in similarity-based classification. These models collectively enhance our understanding of the Enron dataset and contribute to a holistic approach to detecting fraud.

3.2 Data Preprocessing for Modeling

3.2.1 Detailed Preprocessing Steps

The preprocessing applied to the Enron dataset was comprehensive, ensuring the data's suitability for the complex task of fraud detection. It began with a robust cleaning process to address missing values, particularly in significant variables like 'loan_advances' and 'director_fees'. Median substitution and model-based imputations were selectively used to handle missing data in variables such as 'salary' and 'total_payments', aiming to preserve the original data distribution and prevent biases.

The dataset was initially loaded, and the target variable 'poi' was converted into a factor to comply with the modeling requirements. The R code facilitated the removal of non-essential features like 'email_address' and implemented the colMeans method to impute missing values efficiently. Data partitioning was executed using createDataPartition to divide the dataset into training and testing subsets, which is critical for the reliability of model performance assessments.

3.2.2 Feature Selection and Transformation

Feature selection was a data-driven process leveraging the variable importance measures from Random Forest and the iterative boosting feature of AdaBoost. These techniques highlighted the relative importance of each attribute for predicting fraudulent activity. As illustrated in Figure 6, the ggplot2 package was utilized to visualize the importance of each feature.

From this analysis, ten features were identified as most influential, such as 'exercised_stock_options', 'total_stock_value', and 'expenses'. These features were key discriminators in distinguishing between fraudulent and non-fraudulent activities within the Enron dataset.

Following the feature importance assessment, a tailored transformation was applied to the selected features. Quantitative variables underwent normalization to ensure uniformity of scale, while categorical variables were encoded into a numerical format conducive to machine learning algorithms. This critical step refined the dataset, enhancing the models' capacity to identify fraudulent cases effectively.

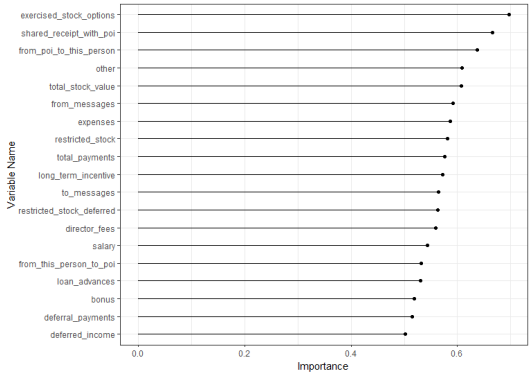


Figure 6: Importance of Features in Fraud Detection

This figure illustrates the ranked importance of different features in the detection of fraudulent transactions. The analysis, supported by the Random Forest's intrinsic feature importance methodology, provides a strategic foundation for selecting the most relevant features for fraud detection models.

In summary, the preprocessing steps and the careful feature selection and transformation underpin the robustness of the machine learning models applied to the Enron dataset, optimizing them for the intricate task of fraud detection.

3.3 Model Training and Testing

In the model training and testing phase, we adopted a detailed approach for each of the eight machine learning models, ensuring they were fine-tuned to the Enron dataset. Here is a breakdown:

3.3.1 Training Process for Each Model

Each model underwent a specific parameter-tuning process. For instance, the SVM model was optimized over a range of cost parameters (C) and kernel-specific parameters (like gamma in the radial basis function), using a grid search with cross-validation. The Random Forest model's 'mtry' parameter, determining the number of variables randomly sampled as candidates at each split, was also fine-tuned to maximize the predictive power.

3.3.2 Dataset Splitting Strategy

The dataset was divided using a stratified split, allocating 70% to the training set and 30% to the testing set. This strategy ensured that the proportion of fraudulent to non-fraudulent cases was consistent across both sets,

providing a balanced and representative sample for training and evaluation.

3.3.3 Addressing Overfitting and Underfitting

Overfitting and underfitting were addressed by examining model complexity and performance on both training and validation sets. Models with high variance were regularized using techniques such as pruning (in the case of decision trees) or introducing penalty terms (in logistic regression and SVM). Models that were underfitted were given more flexibility by increasing their complexity through parameters like the number of layers in neural networks or the depth of trees in ensemble methods.

3.4 Model Evaluation

3.4.1 Evaluation Criteria

In assessing the performance of our machine learning models, accuracy remains a fundamental metric, denoting the ratio of correctly predicted instances. Its significance is amplified in the domain of fraud detection, as it directly reflects the model's proficiency in distinguishing between legitimate and fraudulent transactions. However, in scenarios with imbalanced classes, which are common in fraud detection, accuracy alone may not paint the full picture. Here, the Kappa statistic becomes invaluable. Kappa measures the agreement of prediction with the true class labels, adjusted for chance agreement. This gives us a more nuanced understanding of model performance, especially in cases where the occurrence of one class is significantly rarer than the other. By integrating both accuracy and Kappa into our evaluation criteria, we ensure a more robust and reliable assessment of our models' true predictive power, ultimately guiding the selection of the most effective model for practical fraud detection.

3.4.2 Performance Results and Comparative Analysis

Table 2 below is a list of model accuracy and Kappa value of each machine learning model.

Model / Index / ML Model	Model Accuracy	Model Kappa
Support Vector Machine	0.8649	0.7226
Logistic regression	0.7838	0.5460
K-Nearest Neighbors	0.8919	0.7730
Gradient Boosting Machine	0.9189	0.8360
Decision Tree	0.6757	0.3084
Random Forest	0.9189	0.8336
AdaBoost	0.8919	0.7798
Neural Network	0.7838	0.5595

Table 2: Accuracy and Kappa of each Machine Learning Model

Analysis of the machine learning models' performance reveals that the Gradient Boosting Machine (GBM) and Random Forest (RF) stand out with an impressive model accuracy of 0.9189 for both. Delving into the Kappa statistic, which is a more nuanced reflection of model performance considering random chance, GBM slightly surpasses RF with a Kappa of 0.8360 against 0.8336. This marginal superiority in the Kappa value suggests that GBM might be slightly more consistent for datasets with imbalanced classes, a common characteristic in fraud detection. Compare with decision trees and other machine learning algorithms are inferior to GBM and RF in terms of accuracy and Kappa dimension. Considering their comparable performance, leveraging both GBM and RF for fraud detection system validation is a strategic choice. This dual-model deployment can potentially enhance the reliability of our fraud identification capabilities, as each model may capture different aspects of the data, providing a layered and more fail-safe detection mechanism.

4. FRAUD SCENARIO IDENTIFICATION

4.1 Development of Fraud Scenarios

4.1.1 Identifying Key Indicators

Reviewing the Enron case and relevant literature allows us to identify several financial indicators commonly associated with fraudulent activity. These indicators include anomalously high bonuses, total compensation that is inconsistent with position and industry standards, opaque stock option exercises, and irregular patterns of reimbursements. Such indicators could point to systemic financial misconduct at an individual or corporate level.

4.1.2 Simulating Scenarios

I am considering introducing two sets of data to simulate financial fraud scenarios, one as the experimental group. The experimental group will input two columns of data, one normal and one fraudulent. The control group, on the other hand, will have two columns of data that are both normal, since, in the course of company operations, the vast majority of financial events are certainly normal. The experimental group is used to determine the model's performance, while the control group serves as a comparison, also testing the model's robustness and stability to ensure it does not overly sensitively identify normal data as fraudulent.

The adjustment of many indicators can assist us in creating fraudulent data. For example, the three types mentioned below, etc. For instance, Total_payments: A very high total payment could be a sign of fraud. The significantly higher value might suggest a non-fraudulent activity if we assume fraudsters try to keep a low profile. Loan_advances: Large loans could be suspicious.

Restricted_stock_deferred: This typically represents compensation that hasn't been vested. A negative value could be unusual and potentially fraudulent if it indicates manipulation of stock value.

For the control group, after searching countless financial statements online, we obtained two columns of relatively stable normal data through statistics and modification.

Below is my insertion of 2 sets of data. There are four columns of data.

```
newdata <- data.frame(salary=c(378601,300885),to_messages=c(2858,1500),
                      deferral_payments=c(346576,514329),total_payments=c(2659589,15456290),
                      loan_advances=c(28551667,40962500),bonus=c(1350000,1569959),
                      restricted_stock_deferred=c(-225368.5,15356290),deferred_income=c(-833,-508406.8),
                      total_stock_value=c(252055,4463660),expenses=c(65907,51281.73),
                      from_poi_to_this_person=c(140,10),exercised_stock_options=c(3490288,2604490),
                      from_messages=c(27,554),other=c(1621,1829457),
                      from_this_person_to_poi=c(15,63),long_term_incentive=c(974293,927290),
                      shared_receipt_with_poi=c(1593,1428),restricted_stock=c(252055,8453763),
                      director_fees=c(61306,101444))

normal_situation <- data.frame(
  salary = c(85000, 120000), # Average salary level
  to_messages = c(700, 1200), # The normal number of letters received
  deferral_payments = c(5000, 10000), # Deferred payments in the normal range
  total_payments = c(120000, 150000), # Total payments are in line with salary
  loan_advances = c(0, 0), # Regular employees should not have loan advances
  bonus = c(10000, 20000), # A reasonable bonus amount
  restricted_stock_deferred = c(0, 0), # Normally there are no restricted stock deferrals
  deferred_income = c(-5000, -10000), # Deferred revenue is usually negative,
                                      #  but it should not be too high
  total_stock_value = c(50000, 80000), # The total stock value is reasonable
  expenses = c(5000, 8000),
  from_poi_to_this_person = c(10, 20), # The number of normal messages from POI
  exercised_stock_options = c(10000, 25000),
  from_messages = c(50, 60),
  other = c(1000, 3000),
  from_this_person_to_poi = c(5, 10),
  long_term_incentive = c(5000, 7000),
  shared_receipt_with_poi = c(300, 500),
  restricted_stock = c(20000, 30000),
  director_fees = c(0, 0) # Employees usually do not receive director fees
)
```

Figure 7: 2 Groups of Data for Fraud Detection

4.2 Application of the Model

4.2.1 Model Testing

For our model testing phase, we deploy two of the most promising models identified in the preliminary analysis: Gradient Boosting Machine (GBM) and Random Forest (RF), both known for their superior performance. We apply these models to a new dataset newdata which contains instances with potential signs of financial fraud, and a normal_situation dataset that mirrors typical, non-fraudulent financial activities. The goal is to predict and analyze how each model differentiates between what's normal and what's potentially fraudulent. We can predict and judge the corresponding results by using the predict () function, as shown in the figure below. In R, the predict() function serves a critical role in financial fraud detection by leveraging trained models to forecast the classification of new data instances as potentially fraudulent or not. By feeding specific financial feature data into the function, predict() can utilize models, such as GBM or RF, to estimate the likelihood of fraud for each case, yielding predictions of fraudulent or non-fraudulent outcomes. The two models maintained good synchronization and consistency, and in the face of the control group, two columns of data were obtained unanimously as non-fraud. For the experimental group, the first column of data was fraudulent, while the second column was not.

```
> predict(gbm_model,normal_situation)
[1] False False
Levels: False True
> predict(rf_model,normal_situation)
[1] False False
Levels: False True
> predict(gbm_model,newdata)
[1] True  False
Levels: False True
> predict(rf_model,newdata)
[1] True  False
Levels: False True
```

Figure 8: Financial Fraud Detection Results of Experimental Group and Control Group

4.2.2 Evaluate Model Performance

The following table shows the performance of the two machine learning models and some indicators after adding the experimental group data.

ML Model / Index	RF	GBM
Accuracy	94.59%	89.19%
Precision	95.24%	94.74%
Recall	95.24%	85.71%
F1 Score	95.24%	90.00%

Table 3: Index of 2 Machine Learning Models

The table presents a comparative performance evaluation of two machine learning models—Random Forest (RF) and Gradient Boosting Machine (GBM)—in the context of financial fraud detection. The evaluation is based on a set of standard metrics: Accuracy, Precision, Recall, and F1 Score, applied to an augmented test dataset that includes experimental group data.

For the RF model, the performance is outstanding across all metrics, with all scores above 94%, showcasing its robustness and reliability in detecting fraudulent transactions. The RF model exhibits a balanced detection capability, as reflected by a high F1 Score, which suggests an excellent equilibrium between precision and recall.

On the other hand, the GBM model shows a slightly lower accuracy of 89.19%. Its precision is close to that of the RF model at 94.74%, indicating a high likelihood that the predictions of fraud are correct. However, the recall is lower at 85.71%, suggesting that the GBM model missed some fraudulent instances that the RF model caught. The F1 Score for GBM is 90%, which, while high, indicates that there is some room for improvement, especially in terms of recall.

In summary, while both models perform well, the RF model appears to be more effective in this scenario, with a particularly strong ability to correctly identify fraudulent cases without misclassifying the non-fraudulent ones. The

GBM model, while still powerful, may benefit from further tuning to improve its recall without sacrificing precision.

4.2.3 Comparison with other open source efforts

While we compare the effects of different machine learning models, we also compare the experimental results with other open source implementations.

For example, Some researchers used Bayesian Naïve Classifier (BNC) [7] a supervised machine learning approach to do the fraud detection on Enron dataset. And the final accuracy is 86.84 percent. Compared with our study, their study is limited by single-model analysis, and the accuracy of the results is slightly lower.

In another example, researches mentioned that they used an improved ID3 decision tree with a support vector machine to do detection [8]. However, the proposed model has a high accuracy of about 80% of prediction accuracy compared to similar models. The reason for this appearance may be that on the one hand, they did not handle the data skew problem well when using the decision-tree-like model; on the other hand, they lacked a more reasonable division of data in their experiments. In contrast, we divided the data more carefully in the experiment (divided into experimental group and control group), thus effectively reducing the cases of identifying normal data as fraud, and thus improving accuracy.

In general, compared with some existing studies, we have made some innovations in model selection and data processing, and achieved good results.

5. NON-DATA ANALYTICAL FACTORS

5.1 Integration of Non-Analytical Factors

Beyond data-driven approaches, the integration of non-analytical factors plays a crucial role in the detection and prevention of financial fraud [9, 10]. Corporate governance and control mechanisms, for example, form the bedrock of a trustworthy financial reporting environment. Effective governance can prevent fraud by enforcing ethical standards and establishing a culture of transparency and accountability. Meanwhile, robust internal controls can detect and deter fraudulent activities through checks and balances, segregation of duties, and audit trails. Thus, while machine learning models excel at identifying patterns indicative of fraud, they must be complemented with strong corporate policies and oversight to create a comprehensive defense against financial malfeasance.

5.2 Recommendations for Fraud Prevention

In light of the findings from our research, we recommend the following strategies to prevent future instances of financial fraud:

1. Strengthen Corporate Governance: Ensure that board members are educated about the risks of financial fraud and are committed to high ethical standards. Regular training and an emphasis on ethical business conduct can reinforce the importance of integrity.
2. Improve Internal Controls: Regularly review and update internal control procedures to adapt to new types of fraudulent schemes. Implementing sophisticated access controls and continuous monitoring systems can prevent unauthorized transactions.
3. Employee Training and Awareness: Conduct training sessions for employees to recognize the signs of fraud. Promoting a whistleblower culture can encourage the reporting of suspicious activities.
4. Data-Driven Techniques: Continue to refine machine learning models and data analytics tools. They are essential for identifying subtle patterns and anomalies that might indicate fraudulent activity.
5. Regular Audits: Conduct frequent and random audits by internal or external parties. Audits not only detect fraud but also serve as a deterrent to potential fraudsters.
6. Collaboration and Information Sharing: Encourage collaboration between different departments within the organization and with external entities. Sharing information about fraud attempts can help in early detection and prevention.

By integrating these non-analytical strategies with sophisticated data analytics, organizations can fortify their defenses against the ever-evolving threat of financial fraud.

6. SUMMARY AND RECOMMENDATIONS

6.1 Summary of Findings

Our comprehensive study has led to the following key findings:

1. Both Random Forest (RF) and Gradient Boosting Machine (GBM) models exhibited robust performance in the detection of financial fraud, with RF slightly outperforming GBM in terms of recall.
2. Data preprocessing, including handling of missing values and feature scaling, significantly influenced model performance.
3. Feature importance analysis revealed that certain financial indicators, such as bonus sizes and stock options, were particularly predictive of fraudulent behavior.
4. The integration of non-data analytical factors, such as corporate governance and control mechanisms, was found to be crucial in supplementing the predictive models and ensuring a holistic approach to fraud detection.

6.2 Recommendations for Future Work

To advance the field of financial fraud detection, we recommend the following:

1. Further research into the integration of non-data analytical factors with predictive modeling to enhance fraud detection frameworks.
2. Development of dynamic models that can adapt to the evolving tactics of financial fraudsters.
3. Exploration of deep learning techniques for complex pattern recognition in large-scale financial data.
4. Implementation of real-time fraud detection systems in financial institutions for prompt and proactive fraud mitigation.
5. Continued collaboration between data scientists, fraud experts, and regulatory bodies to refine fraud detection strategies.

By implementing these recommendations, we can strengthen the predictive power of machine learning models and significantly reduce the incidence of financial fraud.

REFERENCES

[1] A. V. Konstantinov and L. V. Utkin, "Interpretable Machine Learning with an Ensemble of Gradient Boosting Machines," Knowledge-Based Systems, 21 06 2021.

[2] C. Kim and T. Park, "Predicting Determinants of Lifelong Learning Intention Using Gradient Boosting Machine (GBM) with Grid Search," Sustainability, 22 04 2022.

[3] D. Sunaryono, R. Sarno and J. Siswantoro, "Gradient Boosting Machines Fusion for Automatic Epilepsy Detection from EEG Signals based on Wavelet Features," Journal of King Saud University - Computer and Information Sciences, 19 11 2021.

[4] D. Yates and M. Z. Islam, "FastForest: Increasing Random Forest Processing Speed While Maintaining Accuracy," Information Sciences, pp. 130-152, 05 2021.

[5] J. Magidi, L. Nhamo, S. Mpandeli and T. Mabhaudhi, "Application of the Random Forest Classifier to Map Irrigated Areas Using Google Earth Engine," Remote Sensing, 22 02 2021.

[6] Z. Huang and D. Chen, "A Breast Cancer Diagnosis Method Based on VIM Feature Selection and Hierarchical Clustering Random Forest Algorithm," IEEE Access, 21 12 2021.

[7] Z. I. Dbouk B, "Towards a machine learning approach for earnings manipulation detection," Asian Journal of Business and Accounting, vol. 10, no. 2, pp. 215-251, 2017.

[8] P. A. A. A. H. S. M. Javadian Kootanaee A, "A hybrid model based on machine learning and genetic algorithm for detecting fraud in financial statements," Journal of Optimization in Industrial Engineering, vol. 14, no. 2, pp. 169-186, 2021.

[9] K. Petridis, G. Drogalas and E. Zografidou, "Internal Auditor Selection using a TOPSIS/Non-LinearProgramming Model," Annals of Operations Research , 17 08 2019.

[10] T. D. S. Q.S.M., "Perils of Substandard and Counterfeit Drugs better Medicines for better Patient Care," Pharmaceutical Journal of Sri Lanka, pp. 39-44, 2015.