Financial Fraud Analytics, Assignment 2

Shihao Liang, UID: 3036196673

November 2023

1 Experimental Result Analysis

Algorithm	Balanced Accuracy
RF	0.622
SVM	**0.661**

Table 1: Results of SVM and RF. The data augmentation is conduct 1,2 or 3 times, and the train size is 18, which is determined by the optimal grid search result.

Data Augmentation	Train Size	Number of Trees	Balanced Accuracy
1	18	100	0.567
2	18	100	**0.622**
3	18	100	0.528
1	9	100	**0.622**
2	9	100	0.589
3	9	100	0.572
2	9	50	0.606
2	9	200	0.583
2	9	500	0.583
2	90	100	0.506

Table 2: Detailed random forest results of the experiment. Conduct grid search for the parameter search.

On the Enron dataset, SVM showed superior performance compared to Random Forest. Detailed results are referred to Table 2 and 1. This could be due to several factors:

- Data Features and Dimensions - The Enron dataset contains many high-dimensional features, where SVM typically outperforms Random Forest in small-sample, high-dimensional data.
- Interactions Between Features - SVM, by building an optimal hyperplane, can better capture complex relationships between features.
- Noise and Imbalance in Data - Although Random Forest can handle imbalanced datasets, SVM may perform better in noisy data.

2 Analysis Through Examples

2.1 Case 1: ALLEN PHILLIP K

Salary and Bonus Allen Phillip K has a salary of $201,955 and a bonus of $4,175,000. These figures are substantial but not necessarily indicative of fraudulent behavior on their own.

Email Communication He had significant email interactions, with 2902 emails received and 2195 sent, including 47 received from a POI and 65 sent to a POI. While the communication volume is high, it might not be uniquely indicative of a POI without further context.

Prediction		Reference	
		FALSE	**TRUE**
Prediction	**FALSE**	24	5
	TRUE	6	4

Table 3: Confusion matrixof the best RF result.

Metric	**Value**
Accuracy	0.7179
95% CI	(0.5513, 0.85)
No Information Rate	0.7692
P-Value [Acc ¿ NIR]	0.8301
Kappa	0.2353
Mcnemar's Test P-Value	1.0000
Sensitivity	0.8000
Specificity	0.4444
Pos Pred Value	0.8276
Neg Pred Value	0.4000
Prevalence	0.7692
Detection Rate	0.6154
Detection Prevalence	0.7436
Balanced Accuracy	0.6222

Table 4: Detailed statistics of the best RF result.

Stock Options and Financial Features Exercised stock options worth $1,729,541 and other financial elements are notable but not conclusively indicative of fraudulent behavior.

2.2 Case 2: BELDEN TIMOTHY N

Salary and Bonus Timothy Belden's salary is $213,999, with a substantially higher bonus of $5,249,999. The disproportionately high bonus might raise red flags.

Email Communication A very high number of emails sent to and received from POIs (228 received from POIs, 108 sent to POIs), which is significantly higher than non-POIs and could indicate closer connections with other POIs.

Deferred Income and Stock Options A deferred income of -$2,334,434 and exercised stock options of $953,136, which might be part of compensation schemes related to fraudulent activities.

2.3 Why SVM is More Suitable

Complexity and Subtlety in Data SVM is likely to correctly classify these individuals by considering the subtle interactions of features like email communication with POIs, financial benefits, and their relative proportions. It effectively separates the high-dimensional space where these nuanced indicators exist.

Random Forest Analysis Random Forest might miss these subtle interactions due to its approach of averaging over multiple decision trees. It may not weigh the disproportionate bonuses and the high level of interaction with POIs as effectively as SVM in the context of identifying fraudulent activities.

Case Comparison In the case of ALLEN PHILLIP K, the Random Forest might not flag him as a POI due to less prominent indicators compared to BELDEN TIMOTHY N. However, SVM's nuanced approach to the high-dimensional feature space can distinguish between these subtleties more effectively, potentially leading to more accurate predictions of POI status.

A Appendices

A.1 Introduction

This report aims to analyze and compare the performance of Random Forest and Support Vector Machine (SVM) in processing the Enron dataset.

A.2 Dataset

The enron.csv dataset includes several features, such as employee salaries, email interactions, stock options, and other financial information related to employees. Some features in the dataset contain missing values, which may impact the choice and performance of the models.

Data Preprocessing. For data 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' (Person of Interest) status.
- **Data Sampling and Augmentation:**
  - Define size parameters for training and test datasets.
  - Augment the 'poi_true_data' for balancing the dataset.
  - Randomly sample a subset of 'poi_false_data' for the test set.
- **Preparing Training and Test Datasets:**
  - Combine augmented true and false POI data for training and test sets.
  - Remove non-numeric columns and any column named 'Unnamed: 0'.
  - Convert the response variable to a factor.

A.3 Model Selection

Random Forest. Random Forest is an ensemble learning method that improves model stability and accuracy by constructing multiple decision trees. Its advantages include:

- Strong performance, suitable for high-dimensional data.
- Ability to assess the importance of features.
- Can balance errors in imbalanced datasets.

However, Random Forest also has downsides, such as potential overfitting on noisy datasets and bias when dealing with attributes with many divisions.

Support Vector Machine (SVM). SVM is an effective classification method that distinguishes different categories by finding the optimal hyperplane. Its advantages include:

- Strong theoretical foundation, suitable for small-sample learning.
- Effectively avoids the curse of dimensionality.
- Efficient "transductive reasoning" simplifies classification and regression problems.

But SVM faces challenges in handling large-scale training samples and difficulties in solving multi-classification problems.