7410 Mini Case Study
Real-life Fraud Detection Scenario

Name: Chen, Xingyi(陈星熠)
UID: xychenn
University Number: 3036198102
Email: xychenn@connect.hku.hk

November 29, 2023

1 Introduction

ENRON dataset[2] is a widely used dataset in the field of machine learning and natural language processing. It is named after the energy company Enron Corporation, which became infamous due to a massive accounting scandal that led to its bankruptcy in 2001. The dataset consists of around 500,000 emails sent by Enron employees, which were made public during the subsequent investigations.

This mini-case study utilizes the ENRON dataset for fraud detection. Due to its large size and diverse content, this study uses a simplified version of the ENRON dataset.

In this study, I first performed Exploratory Data Analysis (EDA) on the dataset and then used the preprocessed data for model training. I trained two models, Random Forest[1] and Neural Network[3], and compared the performance of both. I then tested the Random Forest model, which has better performance among the two, using a fraud scenario dataset developed by myself, and analyzed and summarized the above results.

2 Exploratory Data Analysis

2.1 Summary Description of the Dataset

The dataset[2] contends 22 variables with 20 in numeric type and 2 are character type. So I excluded these 2 character variables, namely poi and email address, and then attempted to find out appropriate candidate variables to do the further analysis. I calculated the deviations of those numeric variables, and 12 of them are distributed in relatively large ranges as shown in Figure 1.

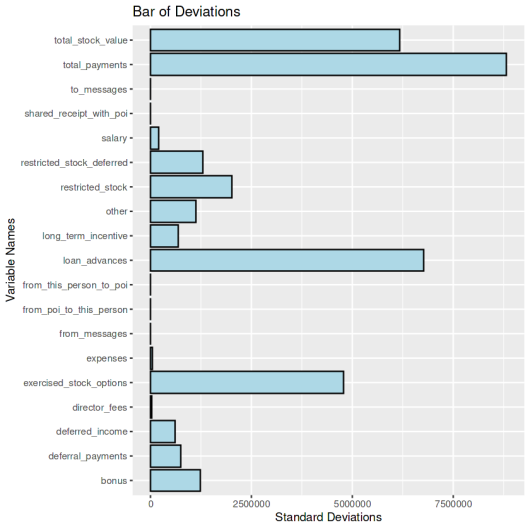

Figure 1: Bar of Deviations

2.2 Univariate Analysis

In this section, I would like to take 3 variables, namely salary, exercised stock options, and bonus, into consideration. I will give my reasons in the following paragraphs.

Firstly, I assume that salary relates to people's rank in their company since a higher position means a higher salary in most cases.

Based on the assumption above, I may further infer that it should present a relatively falling trend of exercised stock options as the salary data goes up. Leaders of the company should exercise their stock options as little as possible since it is wise for them to keep their stock op-

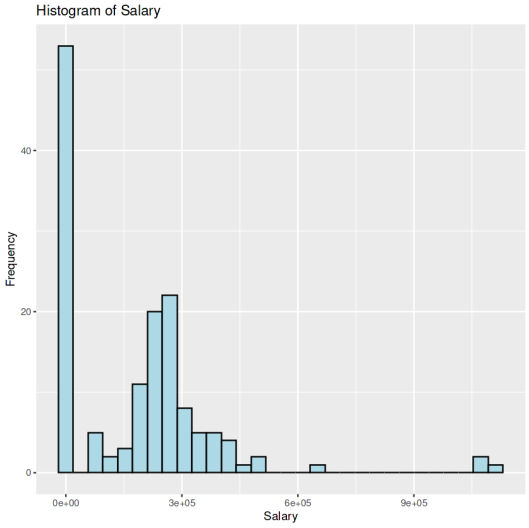

Figure 2: Histogram of Salary

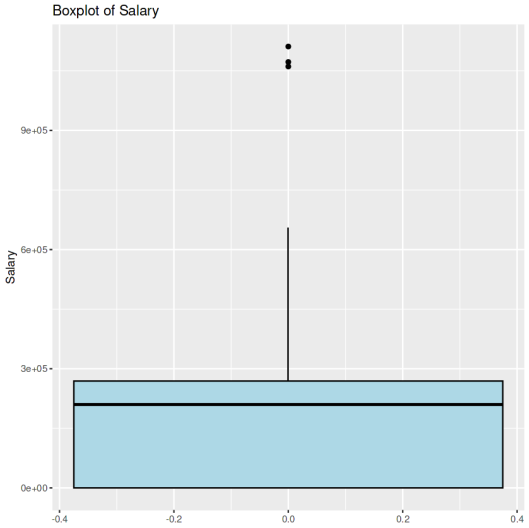

Figure 3: Boxplot of Salary

tions as long as they can if they feel confident about the company's development under most circumstances.

Lastly, bonus data should not be too high compared with the value of its corresponding salary data from my perspective.

Additionally, the standard deviations of these 3 variables are relatively higher than those of others according to the previous section, which is also suspicious. Hence, I choose them to do univariate analysis.

Salary. As shown in Figure 3, there is some outliers' value over 9e05, and the amount of these outliers is really small in Figure 22 So, I infer that these outliers of salary data may come from leaders of the company. I printed their names, and it was a very list of ENRON leaders' names, which validated my previous assumption.

Exercised Stock Options. As shown in Figure 4, there is some outliers' value over 2.5e06. However, it was difficult to identify the dividing line between outliers and maxima from Figure 4, so I drew a histogram with a smaller scale (Figure 5) to find the exact range of outliers, which is above 4e6.

Bonus. As shown in Figure 6, there is some outliers' value over 2e06. I consider these outliers as unusual bonus receivers.

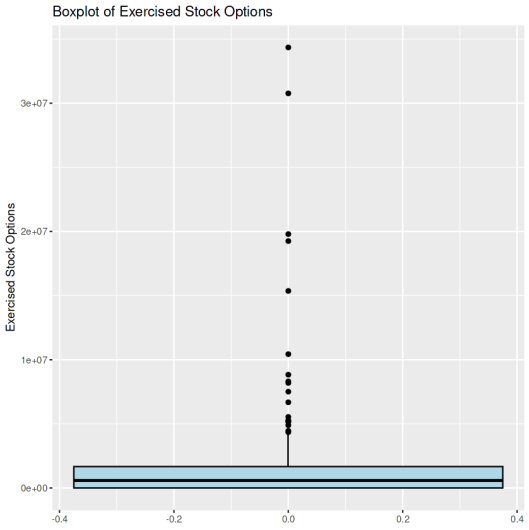

Figure 4: Boxplot of Exercised Stock Options

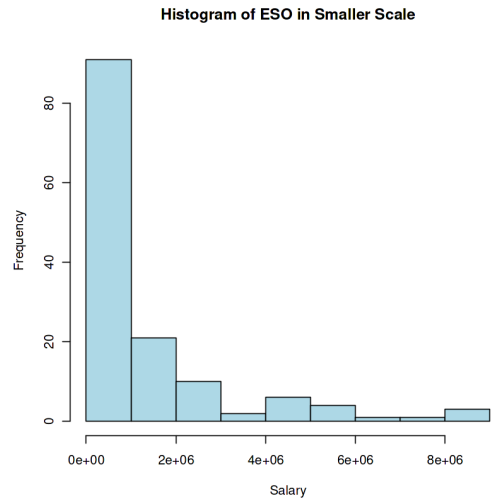

Figure 5: Histogram of ESO in Smaller Scale

2.3 Bivariate Analysis

In this section, I plotted scatter plots for the above three variables two by two to further analyze the relationship

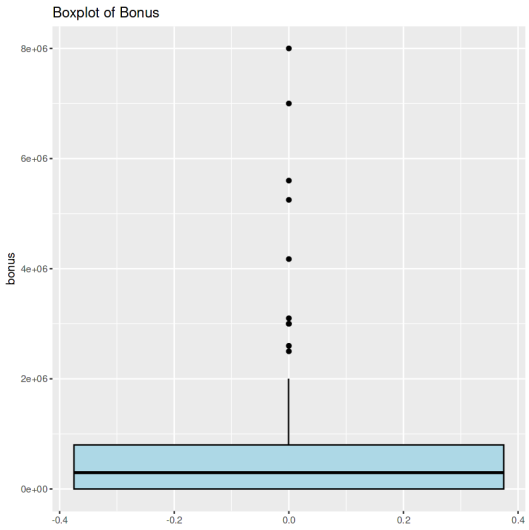

Figure 6: Boxplot of Bonus

between the two variables and to identify bivariate outliers based on my following assumptions.

Association Between Salary and ESO. I have assumed that it should present a relatively falling trend of exercised stock options as the salary data goes up. Leaders of the company should exercise their stock options as little as possible since it is wise for them to keep their stock options as long as they can if they feel confident about the company's development under most circumstances.

However, the scatterplot(Figure 7) shows a number of data points that are not only salary outliers but also ESO(Exercised Stock Options) outliers. It is likely an indication that company leaders are not confident in the company's growth position and may even have engaged in illegal stock transactions.

Association Between Salary and Bonus. Based on the inference that company leaders may sell a large number of their stock, which breaks federal laws, I further assumed that they had accomplices in their employees. Figure 8 shows that there are some people who received bonuses over 10 times their salaries. Hence, I consider these outliers as employees receiving unusual bonuses instead of POI in case some of them are outstanding employees.

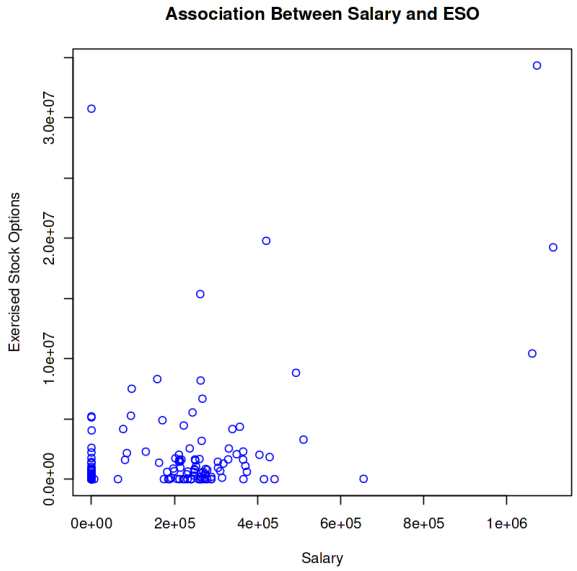

Figure 7: Association Between Salary and ESO

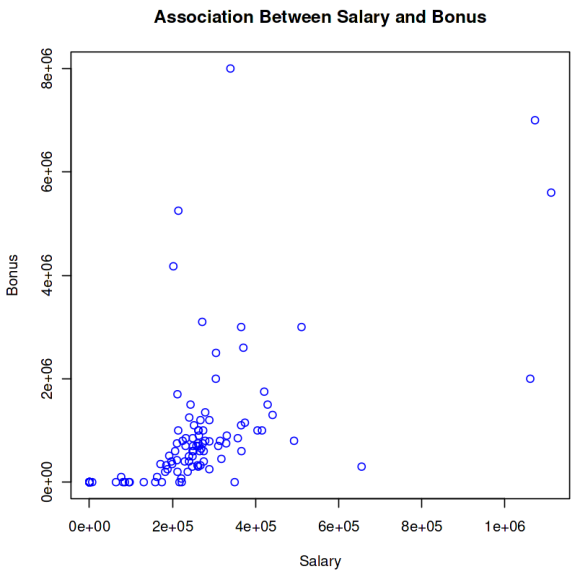

Figure 8: Association Between Salary and Bonus

2.4 Multivariate Analysis

In this section, I plotted correlations among selected variables(Figure 10) where Name and Email Address were removed from. In addition, I transformed poi data from

character to numeric type so that we can see associations between poi data and others.

Figure 10 shows that Salary, Bonus, Total Stock Value, and ESO are significantly more strongly correlated with poi than the other variables. After perusing the dataset meticulously, I found that Total Stock Value is equal to the sum of ESO and Restricted Stock, thus the correlation between Total Stock Value and and ESO is strong. So essentially the strong correlation between Total Stock Value and poi is the strong correlation between ESO and poi. This further validates the validity of the assumptions I made above.

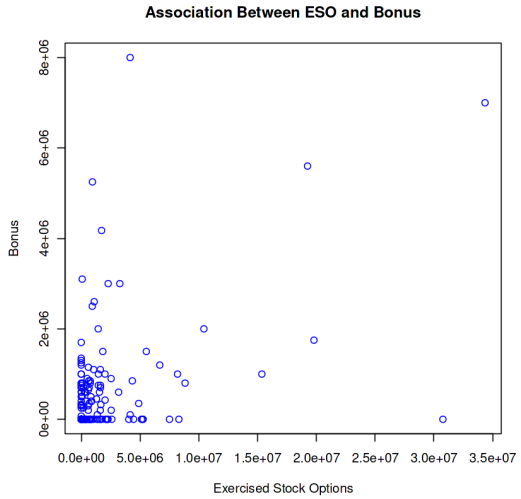

Figure 9: Association Between ESO and Bonus

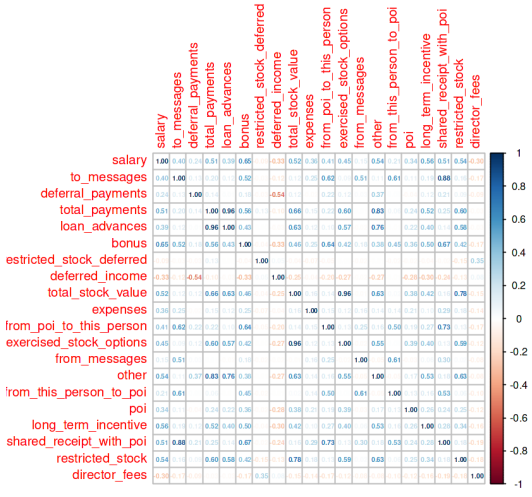

Figure 10: Correlations Among Variables

2.5 Outlier Analysis

In this section, I aggregate univariate outliers, bivariate outliers, and multivariate outliers as a subset of the ENRON dataset and compare their poi values side-by-side.

Figure 11 shows that the more data points that are outliers in more variables at the same time, the greater the probability that their corresponding poi values are true. The joint Salary and ESO variables are used to identify poi are quite effective in the ENRON dataset.

It is worth noting that there is a person in the outlier set named LAVORATO JOHN J whose ESO and Bonus data are outliers but are not poi. I guess that he may be an excellent employee. Data points like this may have an impact on model training, and a decision should be made to either upsample such data or just remove such data based on the parameter capacity of the model.

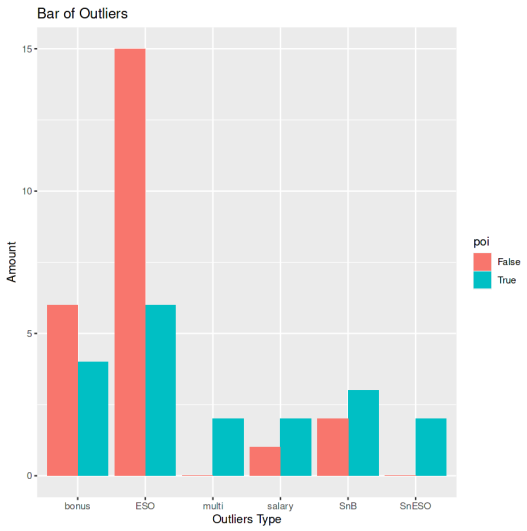

Figure 11: Bar of Outliers

3 Model Building and Validation

3.1 Comparison of Performances of the Two Models

In this section, two supervised learning algorithms, Random Forest[1] and Neural Network[3], are selected to conduct fraud detection modeling on ENRON dataset. Prior to model training, I applied SMOTE (Synthetic Minority Over-Sampling Technique) to enhance the imbalanced dataset, shown as Figure 12 and Figure 13.

I first trained a random forest model, and the final

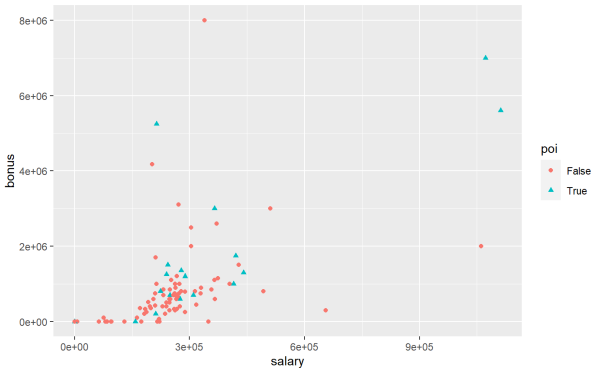

Figure 12: RAW ENRON

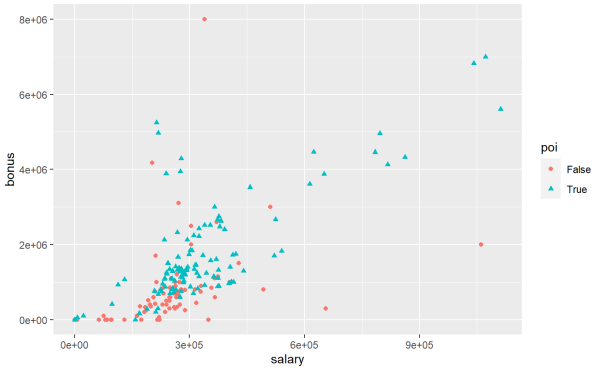

Figure 13: Enhance Imbalanced Dataset

outcome shows an outstanding index of accuracy, which is up to 0.9742. The distribution of this RF model's variable importance is shown in Figure 14 . I then trained a neural network model with 800 epochs, however, within taking more training time, the model didn't get higher performance. The accuracy of the neural network is just 0.9142.

There is a clear performance gap between random forest and neural network, as shown in Figure 16 and Figure 17. Green dots indicate data points that are consistent with the ground truth, and the red ones indicate the wrong prediction. We can see that random forest model did a better job in this task.

3.2 Result Analysis

According to the experiment results above, I then further discuss the shortage and advantage of deploying these two models in real-life situations. In this experiment, the training time of the two models is short and the models' sizes are small due to the small amount of model parameters. So it's possible to run a random forest model or

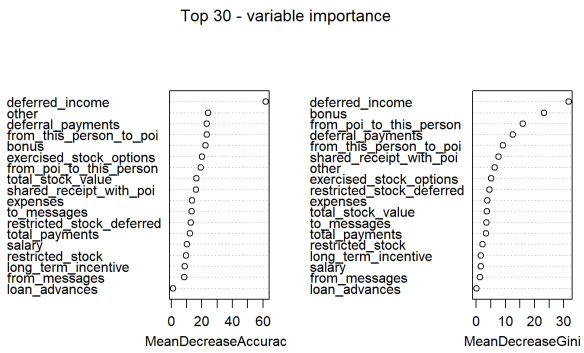

Figure 14: RF Variable Importance

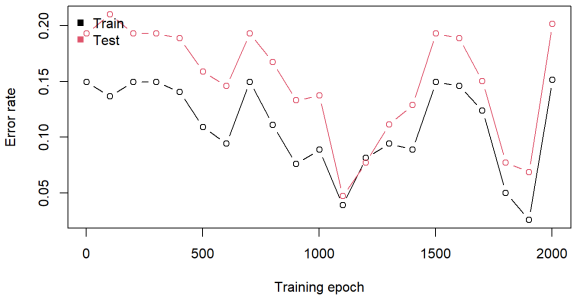

Figure 15: Process of NNet Training

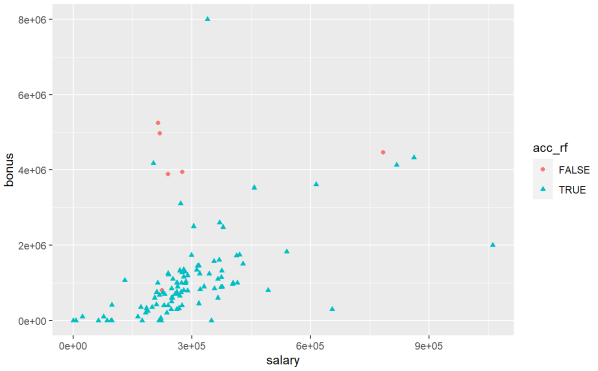

Figure 16: Accuracy of RandomForest

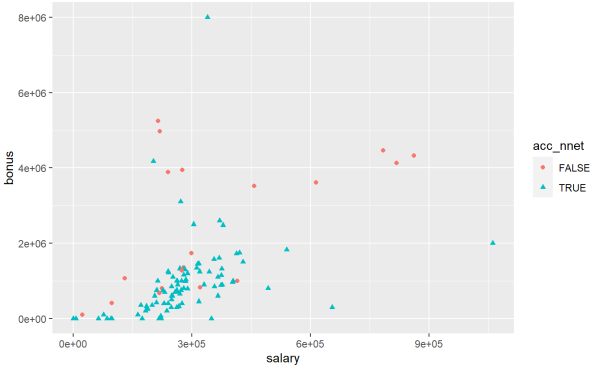

Figure 17: Accuracy of Neural Network

Table 1: Prediction vs Groundtruth

		RF Prediction	
		False	True
GT	False	406	44
	True	11	49

neural network model online, and they could be trained online for several certain trading cases. However, fewer model parameters means weaker learning ability. The random forest model can only handle some simple tasks while it's risky to apply random forest models to some complex fraud-detecting tasks. Neural network models although perform weaker than random forest models in this experiment, can improve their learning ability by extending more neural network layers. In general, it's more efficient to deploy random forest models in simple fraud detection, and deploying deeper neural network models on the complex fraud detection task could be more reliable.

4 Fraud Scenario Identification

4.1 Permutation of Fraud Scenarios

4.2 Test Financial Fraud Scenario Definition

Insider Trading

4.3 Test Results of RF Model

4.4 Result Analysis

5 Summary and Recommendation

5.1 Data-based Analysis

5.2 Non-data-based Analysis

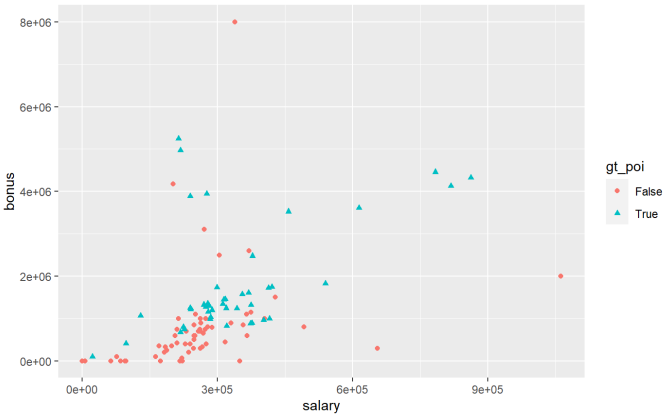

Figure 18: Groundtruth

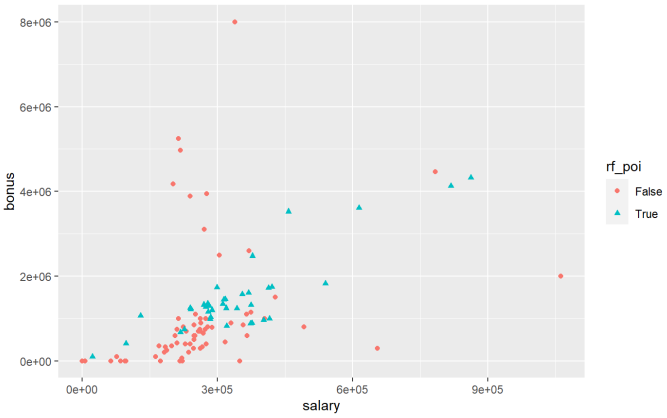

Figure 19: Prediction of RandomForest

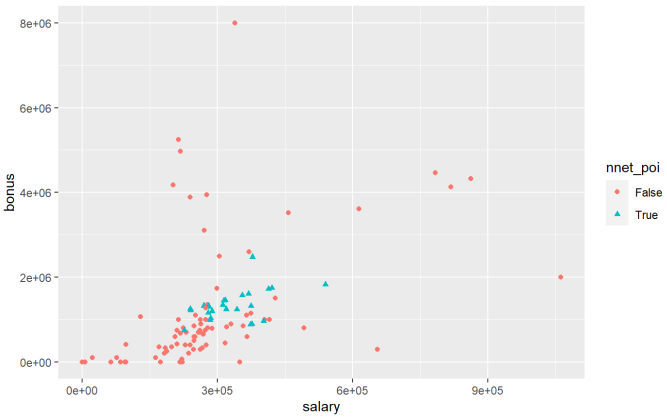

Figure 20: Prediction of Neural Network

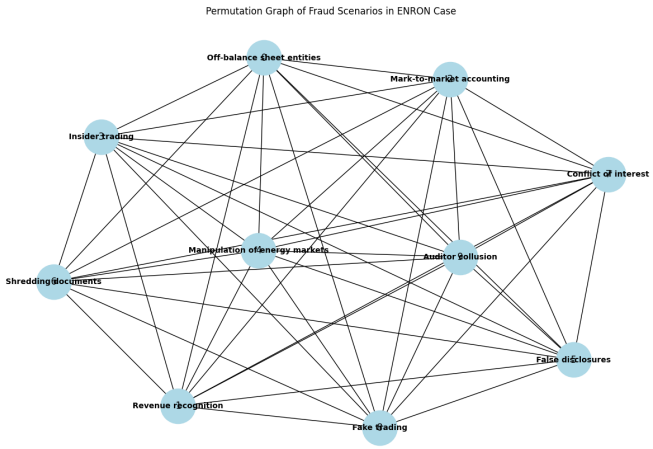

Figure 21: Permutation Graph of Fraud Scenarios

References

[1] Breiman, L. Random forests. *Machine Learning 45*, 1 (Oct 2001), 5–32.

[2] Klimt, B., and Yang, Y. The enron corpus: A new dataset for email classification research. In *Machine Learning: ECML 2004* (Berlin, Heidelberg, 2004), J.-F. Boulicaut, F. Esposito, F. Giannotti, and D. Pedreschi, Eds., Springer Berlin Heidelberg, pp. 217–226.

[3] McCulloch, W. S., and Pitts, W. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics 5*, 4 (Dec 1943), 115–133.

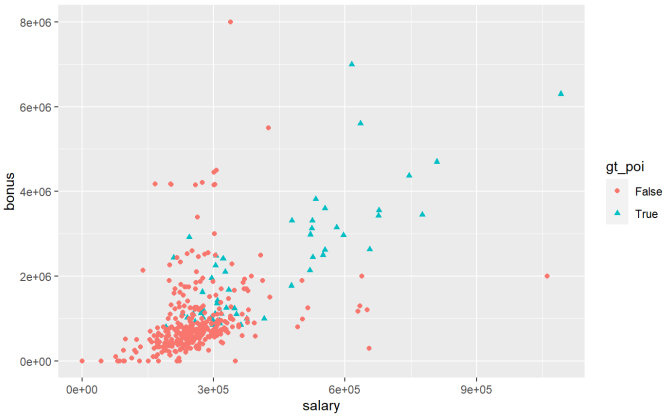

Figure 22: Groundtruth of Testset

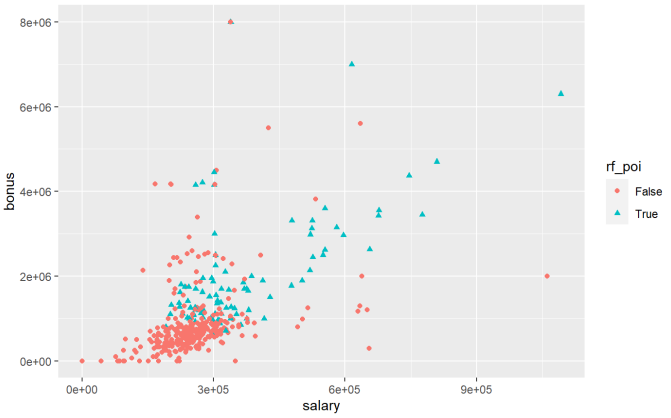

Figure 23: Prediction Results of RF Model on Testset

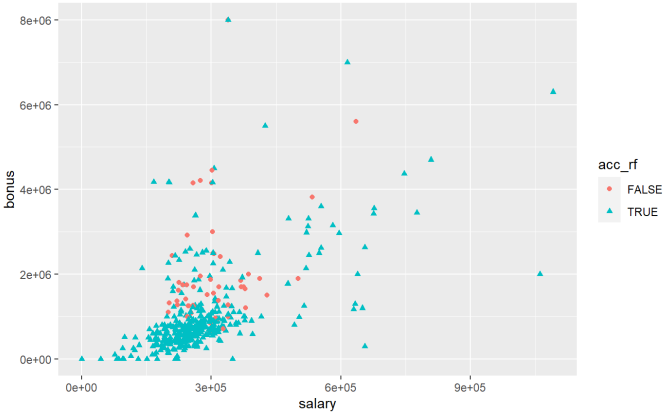

Figure 24: Prediction Accuracy of RF Model on Testset

6 Appendix

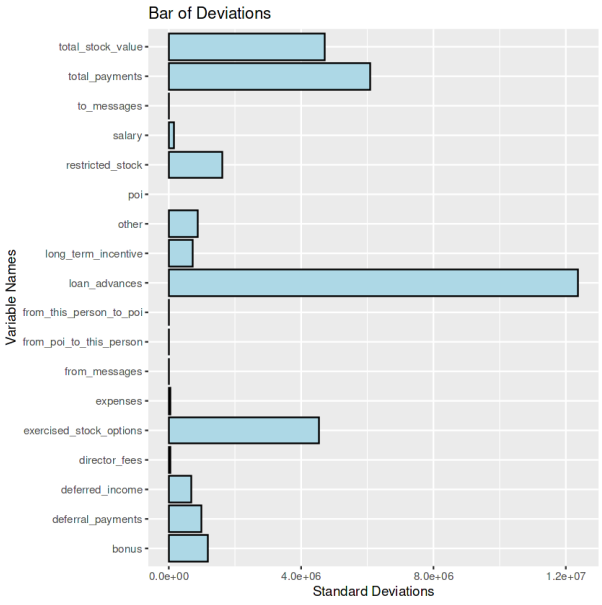

Figure 25: Deviation Distribution of Test Dataset

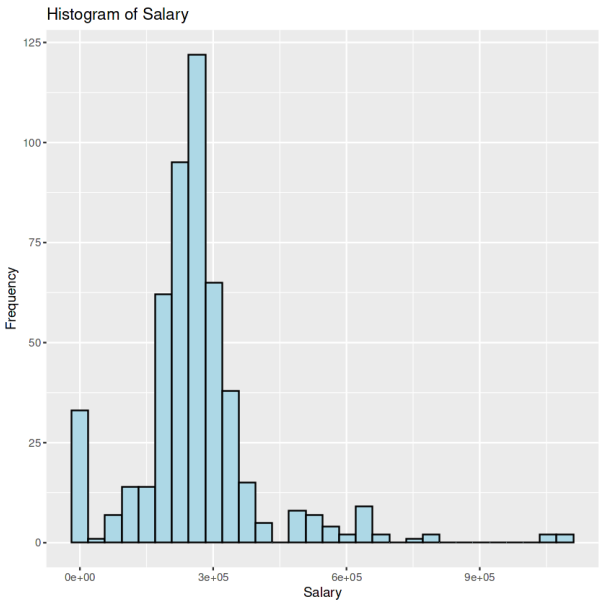

Figure 26: Histogram of Salary on Testset

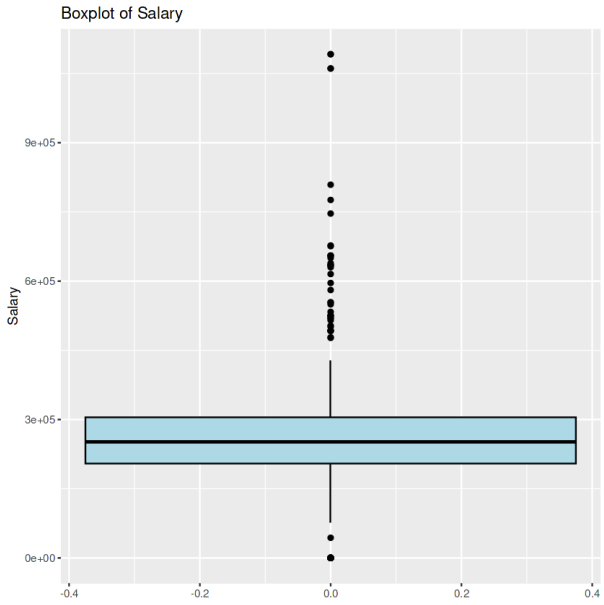

Figure 27: Boxplot of Salary on Testset

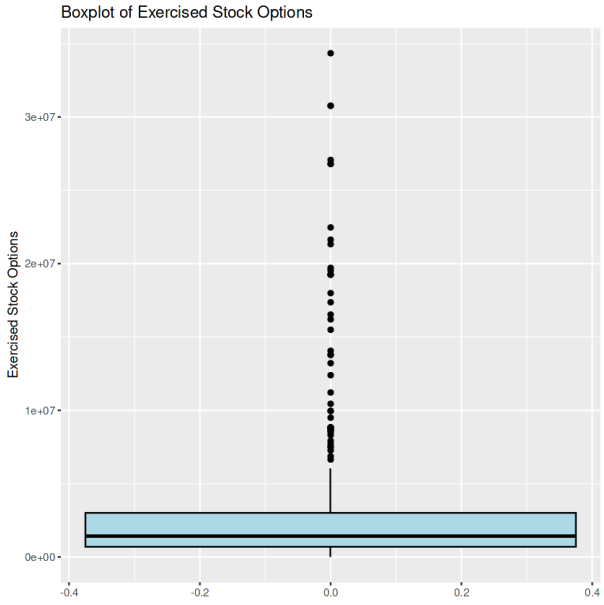

Figure 28: Boxplot of Exercised Stock Options on Testset

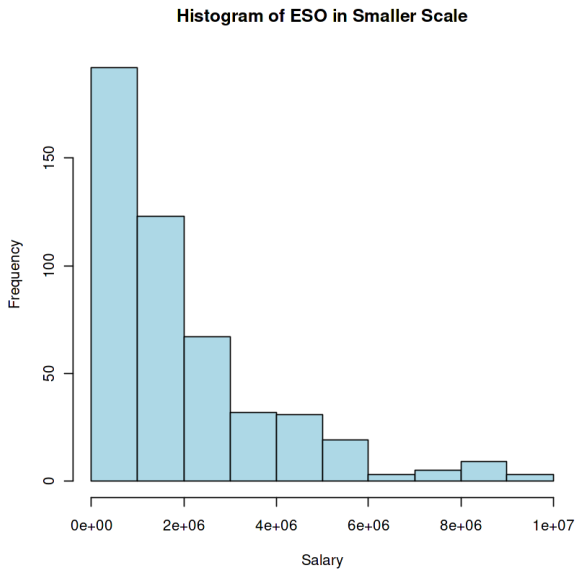

Figure 29: Histogram of ESO in Smaller Scale on Testset

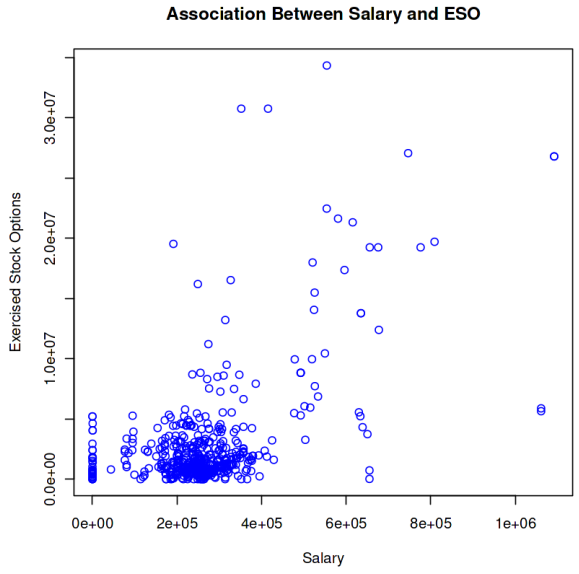

Figure 31: Association Between Salary and ESO on Testset

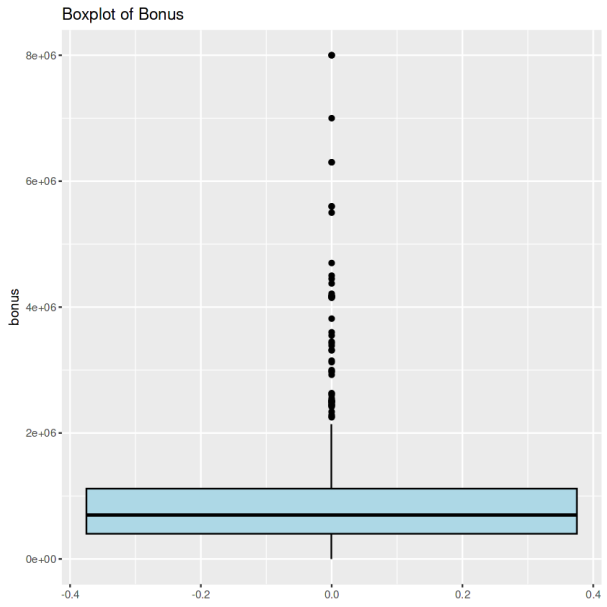

Figure 30: Boxplot of Bonus on Testset

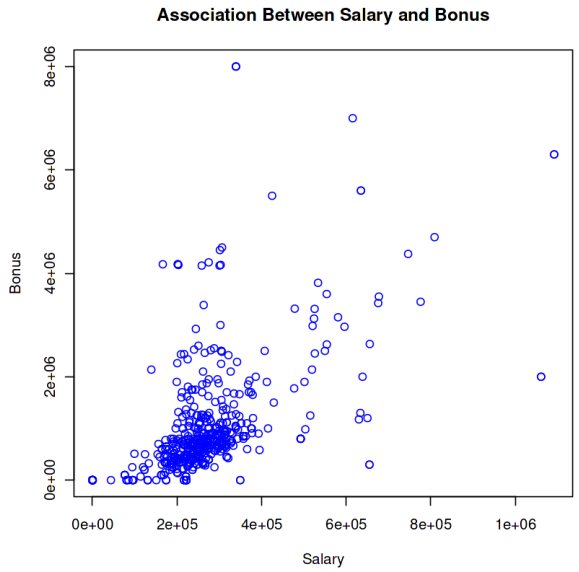

Figure 32: Association Between Salary and Bonus on Testset

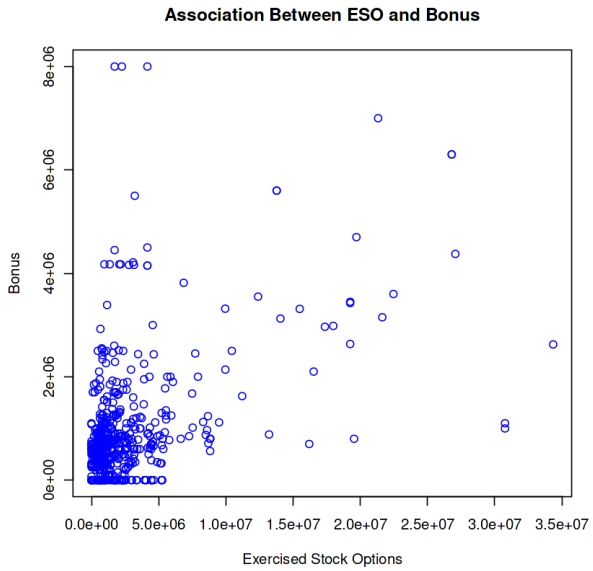

Figure 33: Association Between ESO and Bonus on Testset

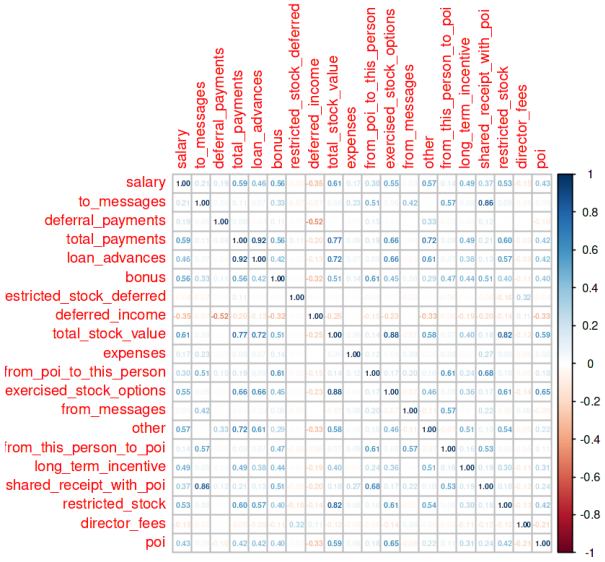

Figure 34: Correlations Among Variables on Testset

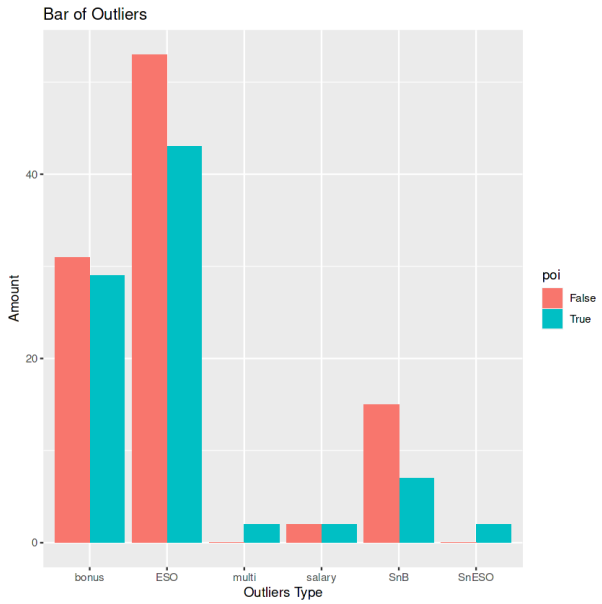

Figure 35: Bar of Outliers on Testset