Enron dataset description based on EDA result

Name: Zheming Kang; UID: 3036195746;

1. **Background Information**

Enron dataset is from enron case, which is a famous financial fraud case. From Britannica, Enron company wrote unrealized future gains into current income statements and abused SPE(special purpose entities) distribution to hide loss.

1. **Summary description of the dataset**

**The dataset consists of 22 variables and 146 data points, which means that the dataset provides us 22 features and 146 people from Enron Company. Among the 22 variables, there are 3 classes containing characters and 19 classes holding numbers. X, email address and poi are the 3 characteristic classes which present name, email address and the status of if the person is a Person of Interest respectively. There are 128 FALSE pois and 18 TRUE pois, which means 87.67% and 12.33% of the dataset. Figure 1 presents details of the dataset.**

文本

描述已自动生成图片包含 日历

描述已自动生成

**Figure 1. summary & structure of dataset**

**We can distinguish the attributes by poi status. The variables present the information of the person. Salary, bonus, loan advances, total stock value, expenses, other, long term incentive, total payment, exercised stock options, director fees and restricted stock describe the individual financial status. Deferral payments, restricted stock deferred, and deferred income provide the information of delayed financial statements. To messages, from poi to this person, from messages, from this person to poi, and shared receipt with poi counts the amount of messages transmitted from the person to others.**

1. **Univariate Analysis**

**After step1:** Distinguish Attributes, we can choose some attributes that may have potential relationships to analyse. For this part, I split the attributes into 3 groups and analysis them respectively.

The first group contains “from poi to this person”, and “from this person to poi”. I group them for their close relationship to poi. The second group contains salary, bonus, total stock value, and expenses since they can show the financial information of the person. As the Enron case was caused by counting delayed income and hiding current loss, I choose the third group attributes which are deferral payments, income, and restricted stock. Figure 2 is the histogram of total stock value. To get rid of the influence of outliers, I plot the attributes in a reasonable range, which reflects a smooth distribution just like in figure 2. Then, I plot the distribution of poi using bar chart. We can see the distribution of two types of people: poi and non-poi in figure 3. Finally, I plot the box plot to see the distribution of personal financial attributes and defer-relative attributes.

图表

描述已自动生成 图表, 直方图

描述已自动生成

Figure 2. distribution of total stock value/ value < 5e6

图表, 条形图

描述已自动生成 图表, 直方图

描述已自动生成

**Figure 3. distribution by poi and box plot of salary**

1. **Bi-/Multi-variate analysis**

**Knowing the distribution of single attributes, we should move to bi-variate analysis. I plot the potentially useful attributes with poi feature and find “from poi to this person”, “total stock value”, “deferral payments”, “deferred income”, and “restricted stock deferred” have different distributions in poi and non-poi. Other attributes reflect similar distributions so we won’t keep them in multi-variate analysis.**

**Locking on these attributes, we could compute the correlation matrix and plot the matrix in heatmap. Figure 4 shows the correlation matrix. From the matrix, we could see the deferred income is highly negatively related with deferred payments, total stock value and restricted stock deferred. What’s more, the deferred income is slightly positive related to the “from poi to this person”. An interesting fact is, the total stock value is negatively related to from poi to this person. From the information above, we could infer that if a person who contains low stock, he might receive many emails from poi. We could also infer that if a person need to pay much back, if his deferred income is highly negative, he might have low total stock value, restricted stock deferred and deferral payment.**

图片包含 表格

描述已自动生成图表

描述已自动生成

**Figure 4. correlation matrix**

1. **Missing data/ Outlier analysis**

**I counted the missing value distribution of various indicators, and the image is as shown in the figure 5.**

图表, 直方图

描述已自动生成

**Figure 5. distribution of missing value**

**For u**nivariate outliers, **I compute the IQR of salary as an instance. In figure 5 we could see the iqr is 100301. Then, let’s assume threshold is 1.5\*iqr, and find all outliers.**

表格

描述已自动生成

**Figure 6. IQR and outliers of salary**

**For Bi-/Multi-varite outliers, we could find them by histogram in session 4. From figure 2 left part, we could see the outlier in the right hand side. Actually the outlier is data “total” in salary. In figure 3 right hand side, we could see the outlier from the box plot. The outlier is far away from the data piece.**

**There are a lot of figures when I doing EDA, the figures above are just examples.**

1. **References**

[Enron scandal | Summary, Explained, History, & Facts | Britannica](https://www.britannica.com/event/Enron-scandal)

[Enron Person of Interest Dataset (kaggle.com)](https://www.kaggle.com/datasets/sagarnildass/enron-person-of-interest-dataset)