DATA SCIENCE IN PRACTICE

(MGT-415)

Fraud Detection using Emails Dataset

Authors:
Firas Kanoun
Nourchene Ben Romdhane
Sami Ben Hassen
Zeineb Sahnoun

Supervisor:
Ph.D. Christopher Bruffaerts

December 11, 2019

Contents

1 Introduction

1.1 Background

Enron Corporation was an American energy and services company based in Houston, Texas. Before its bankruptcy on December 3, 2001, Enron employed approximately 29,000 staff and was a major electricity and natural gas company, with claimed revenues of nearly $101 billion during 2000.

At the end of 2001, it was revealed that Enron's reported financial condition was sustained by institutionalized, systematic, and creatively planned accounting fraud, known since as the Enron scandal.
In addition to being the largest bankruptcy reorganization in American history at that time, Enron was cited as the biggest audit failure.

As was later discovered, many of Enron's recorded assets and profits were inflated or even wholly fraudulent and nonexistent. Enron used a variety of deceptive, bewildering, and fraudulent accounting practices and tactics to cover its fraud in reporting Enron's financial information.
Special Purpose Entities were created to mask significant liabilities from Enron's financial statements. These entities made Enron seem more profitable than it actually was, and created a dangerous spiral in which, each quarter, corporate officers would have to perform more and more financial deception to create the illusion of billions of dollars in profit while the company was actually losing money. This practice increased their stock price to new levels, at which point the executives began to work on insider information and trade millions of dollars' worth of Enron stock.

Enron has since become a well-known example of willful corporate fraud and corruption. The scandal also brought into question the accounting practices and activities of many corporations in the United States. In addition, it caused the dissolution of Arthur Andersen, which at the time was one of the "Big Five" - the world's foremost accounting firms.

Figure 1: Enron Corp. logo

1.2 Problem statement in business

Fraud is a billion-dollar business and it is increasing every year. The PwC global economic crime survey of 2018 found that half (49 percent) of the 7,200 companies they surveyed had experienced fraud of some kind. This is an increase from the 2016 study in which slightly more than a third of organizations surveyed (36%) had experienced economic crime.
A typical organization loses five percent of its annual revenue to fraud, with a median loss of $160,000. Frauds committed by owners and executives are more than nine times as costly as employee fraud. Fraud can also lead to bankruptcy as was the case for Enron which shows at which extent the matter is serious and must be detected at very early stages.

1.3 Problem statement in analysis

The Enron Corpus is a large database of over 600,000 emails generated by 158 employees of the Enron Corporation and acquired by the Federal Energy Regulatory Commission during its investigation after the company's collapse.

At the conclusion of the investigation on Enron bankruptcy, the emails and information collected were deemed to be in the public domain, to be used for historical research and academic purposes.

A copy of the email database was subsequently purchased for $10,000 by Andrew McCallum, a computer scientist at the University of Massachusetts Amherst. He released this copy to researchers, providing a trove of data that has been used for studies on social networking and computer analysis of language.

The corpus is unique because it is one of the only publicly available mass collections of real emails available for study, as such collections are typically because of privacy concerns. We are using the 2015 version of the data publicly available on the CMU website[3].

The main purpose of our project is to use emails exchanged inside Enron to detect fraud but also fraudulent actors through sentiment analysis of emails sent and received by employees. We also use network analysis to study relationships between employees and try to find suspicious relationships.

Aiming at adding a financial sense to our analysis, we also added financial data to our models. This data represented the different payments and bonuses to many of the Enron employees before the scandal. A pdf containing this information is available on findlaw website.[4]

1.4 Stakeholders

We can imagine that Enron company came to us as data scientists for advice before its bankruptcy. Another scenario would be the Investigation team working on the lawsuit coming to us with the data to try to detect implicated employees in the fraudulent activities of Enron.

Several parties can benefit directly or indirectly from the project. The party for which the project is the most profitable is obviously the company, by being able to detect fraudulent operators among its employees and avoiding tremendous incurred financial loss, bankruptcy, loss of clients due to bad reputation.

- The employees who will have to agree to sharing content of their professional email exchanges.
- The IT department of the company will have to put the detection model in place.
- Clients of the company: a clear no fraud policy will in theory attract more clients and give a sustained brand strength.
- The shareholders of Enron: after the bankruptcy, they received limited returns in lawsuits, despite losing billions in pensions and stock prices. Putting a detection model in place will make shareholders more confident when investing in Enron.
- Our consulting company will also benefit from this solution by selling this product to the company/investigation structure as mentioned above.

2 Data fetching and cleaning

2.1 Emails dataset

The data used to cope with the problematic explained above, are under the form of a dataset containing 543'446 emails. Since the size of the data set is to big, we will not be providing it in the submission, but you can download it from [3].

The Data-Set is represented as a list of folders. Each folder contains the Mail-Box of an employee. Therefore each folder contains a number of sub-folders representing for example inbox, outbox... We managed to compile everything in a nice pandas dataframe which helped us in our data exploration. Each email entry is detailed with 19 attributes as shown in Figure 2.

Attribute	Description
file	Full path name of the file
Message-ID	Unique identifier for the email
Date	Date at which the email was sent
From	Sender's email address
To	Receiver's email address
Subject	Email's subject
Mime-Version	Mime version from email header
Content-Type	Type of text content from email header
Content-Transfer-Encoding	Encoding used for the smtp protocol
X-From	Name of the sender as it was defined
X-To	Name of the receiver as it was defined
X-cc	Name of the person in cc field
X-bcc	Name of the person in bcc field
X-Folder	Path where the email was stored on the database
X-Origin	Nametag of the email sender
X-FileName	Name of the file containing the email on the database
content	Text content of the email
user	Nametag of the sender again
subfolder	Folder where the email was stored in the mailbox

Figure 2: Initial attributes description

First thing we did was to get rid of duplicate emails. This enabled us to set Message-ID as the index. We also noticed that the file, Mime-Version, Content-Type, Content-Transfer-Encoding had only one or two possible values and were useless for our analysis so we just dropped those columns.

Another important thing, when dealing with NAN missing values we noticed that if we just dropped all of them it will result in loss of a huge portion of our dataset.

We decided then to drop NAN values in columns which are critical to our analysis: Subject, content, From, To.

Attribute	Description
Message-ID	Unique identifier for the email
Date	Date at which the email was sent
From	Sender's email address
To	Receiver's email address
Subject	Email's subject
X-From	Name of the sender as it was defined
X-To	Name of the receiver as it was defined
X-cc	Name of the person in cc field
X-bcc	Name of the person in bcc field
X-Folder	Path where the email was stored on the database
X-Origin	Nametag of the email sender
X-FileName	Name of the file containing the email on the database
content	Text content of the email
user	Nametag of the sender again
subfolder	Folder where the email was stored in the mailbox

Figure 3: Attributes after data cleaning

2.2 Financial dataset

This part showed us how hard it can be to fetch and compile clean usable data. Since the scandal, a lot of the company's and employees' financial information were made public. We used a pdf of this information available on findlaw[4].

We first started by scrapping the pdf document to extract financial information about employees. Since the scrapping does not work with the same parametres for all 4 pages of the document, we were obliged to scrap each page separately by tuning the parameters.

After scrapping the pdf pages, we organized them in a data frame which has the attributes described in figure 4.

Attribute	Description
Name	Employee name
Salary	Employee salary
Bonus	Employee bonus
Long Term Incentive	Long Term Incentive received
Deferred Income	Employee Deferred Income
Deferral Payments	Employee Deferral Payments
Loan Advances	Employee Loan Advances
Other	Other received payments
Expenses	Payments for employee expenses
Director Fees	Director fee received
Total payments	Total payments perceived by the employee
Exercised Stock options	Exercised Stock options of the employee
Restricted Stock	Restricted Stock of the employee
Total Stock Value	Total Stock Value of the employee

Figure 4: Attributes of financial data

3 Exploratory Data Analysis

3.1 Time is money

In this part of the project we explored different times people tend to send emails inside the company and showed how valuable time in e-mails can be in finding information about the company.

From the yearly graph (Fig.5), we can see that the bankruptcy happened at the very beginning of the year of 2002 since the number of emails drastically decreased. If it had been towards the end we would have had more emails exchanged that year. Checking with Wikipedia[1] we find that in ”Dec. 2, 2001: Enron files for Chapter 11 bankruptcy protection”.

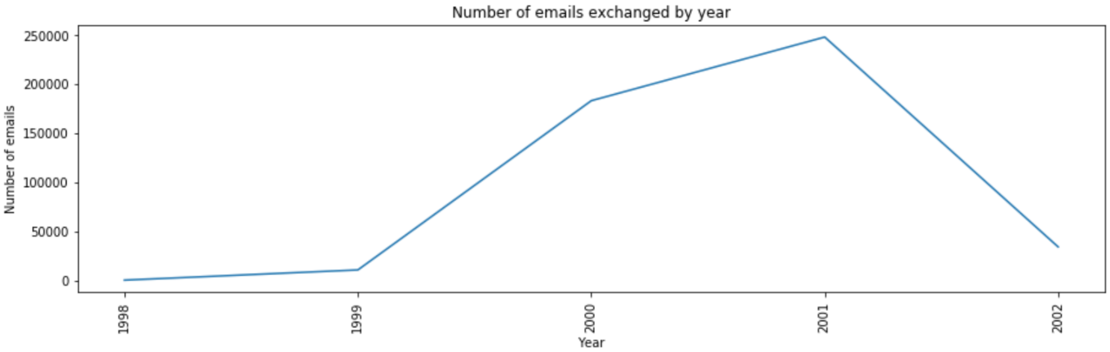


Figure 5: Number of emails exchanged by year.

Employees at this company reach their highest productivity on Tuesdays and Wednesdays as seen in Fig.6. This is in accord with what most studies show [5]. We can also safely assume that employees don’t work on weekends since there is very little activity on Saturdays and Sundays.

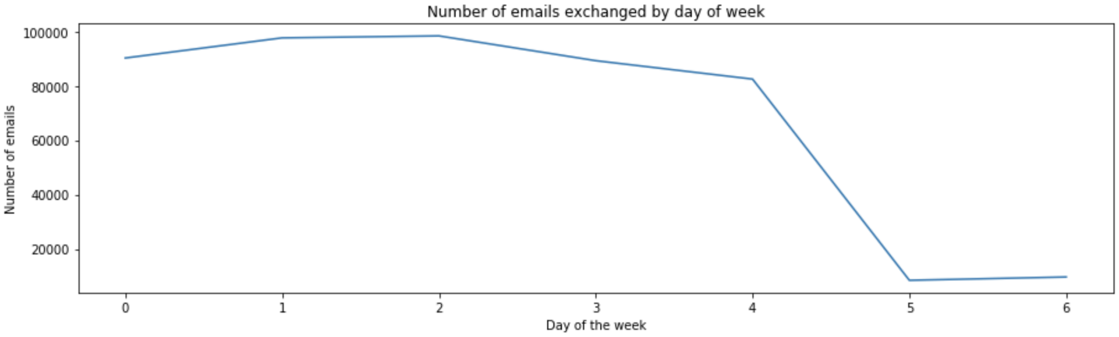


Figure 6: Number of emails exchanged according to the day of the week.

When looking at the number of emails sent by hour (Fig.7) we get the exact schedule in the day of the employees. We can confidently say that they start their day at 8AM and finish at 6PM with a one hour break at noon. They are quite productive at 10AM and reach their highest productivity at 4PM before people start leaving.

This doesn't apply to only ENRON employees though. Studies that have explored productivity in different countries around the world have shown similar productivity patterns[7].

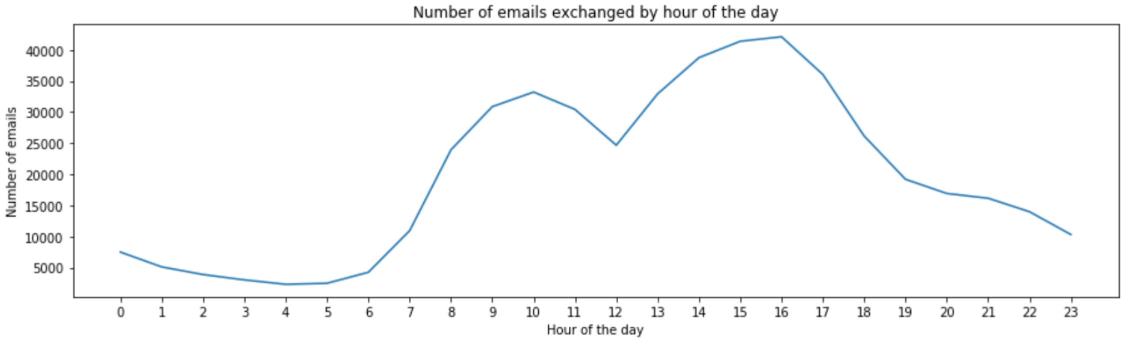


Figure 7: Number of emails exchanged by hour of the day.

3.2 People who sent most emails

After analyzing the pattern of emails sent for each period of time. We wanted to actually see the contents of the emails and who are the most active people within the company. The results are depicted in figure8.

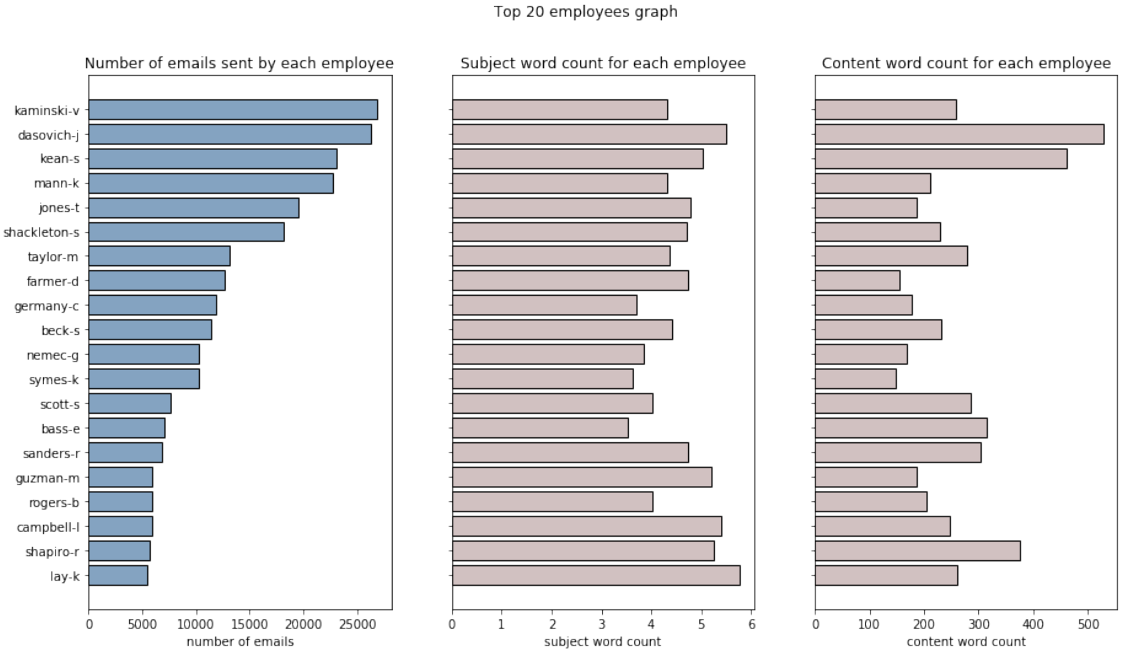


Figure 8: Top 20 users at ENRON

The most active employee was Vincent Julian Kaminski who worked as the Managing Director for Research. In this capacity he led a team of approximately fifty analysts who developed quantitative models to support energy trading

Next is Jeff Dasovich who was the Government Relation Executive.

3.3 What do the emails say?

3.3.1 In the subject

Taking a deeper look to examine the actual content of all these emails and removing some obvious stop words (Fig.9)

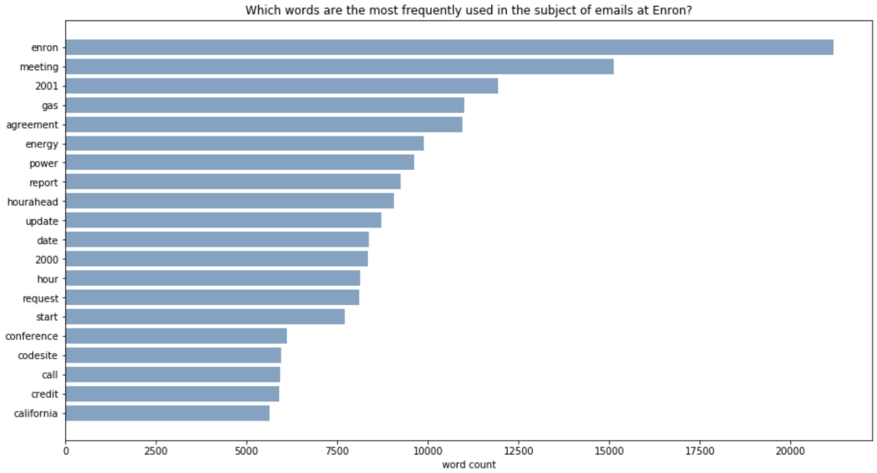


Figure 9: TOP Words used in the subject at ENRON

We find out that the most used words within this company are usually part of the following categories:

- Meetings and time updates : report, date, hour, ahead, codesite, meeting, conference, start, request, call...
- The field the company is working in: gaz, power, energy...
- The years during which the company has been more active: 2000 and 2001.

3.3.2 In the content

The ones at the top is the name of the company and the years it has been more active in.

- Words like gaz, power, energy gives us an idea about what the industry of the company.
- Words like please,time,information ... are also frequent. This means that most of the content exchanged in the company is formal and concerns meetings and time updates.
- We can also remark the presence of the word "California" which we can explain by the secret role Enron played in what has been called the " California Electricity Crisis" between 2000 and 2001 where demand-supply gap were created

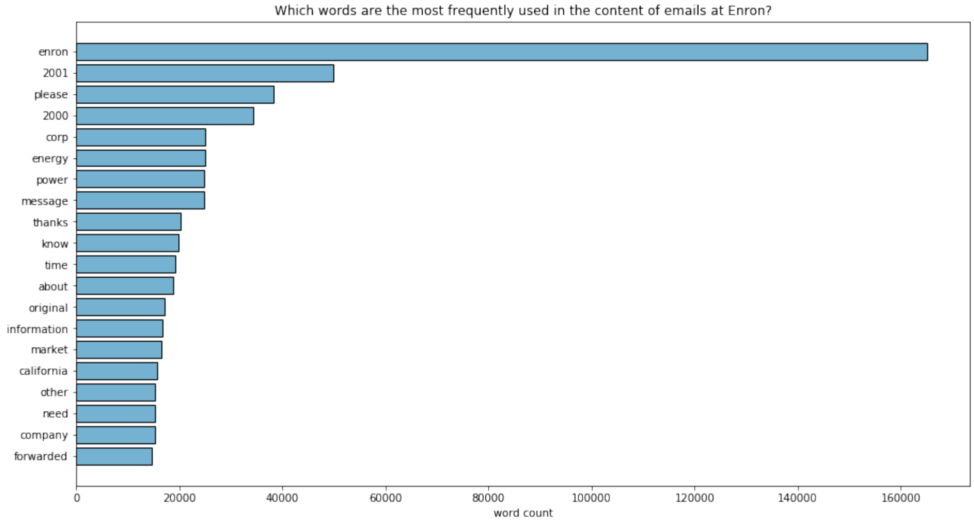


Figure 10: TOP Words used in the content of the emails at ENRON

by energy companies, mainly Enron, to create an artificial shortage. Energy traders took power plants offline for maintenance in days of peak demand to increase the price.[8]

4 Network Analysis

4.1 The graph

We create a graph having as nodes the name of employees with edges representing the connection between them. We will create an index between 0 and 1 that represents the strength of the connection between the nodes. This will allow us to see which people are interacting with who and therefore try to separate the several departments and entities inside the firm.

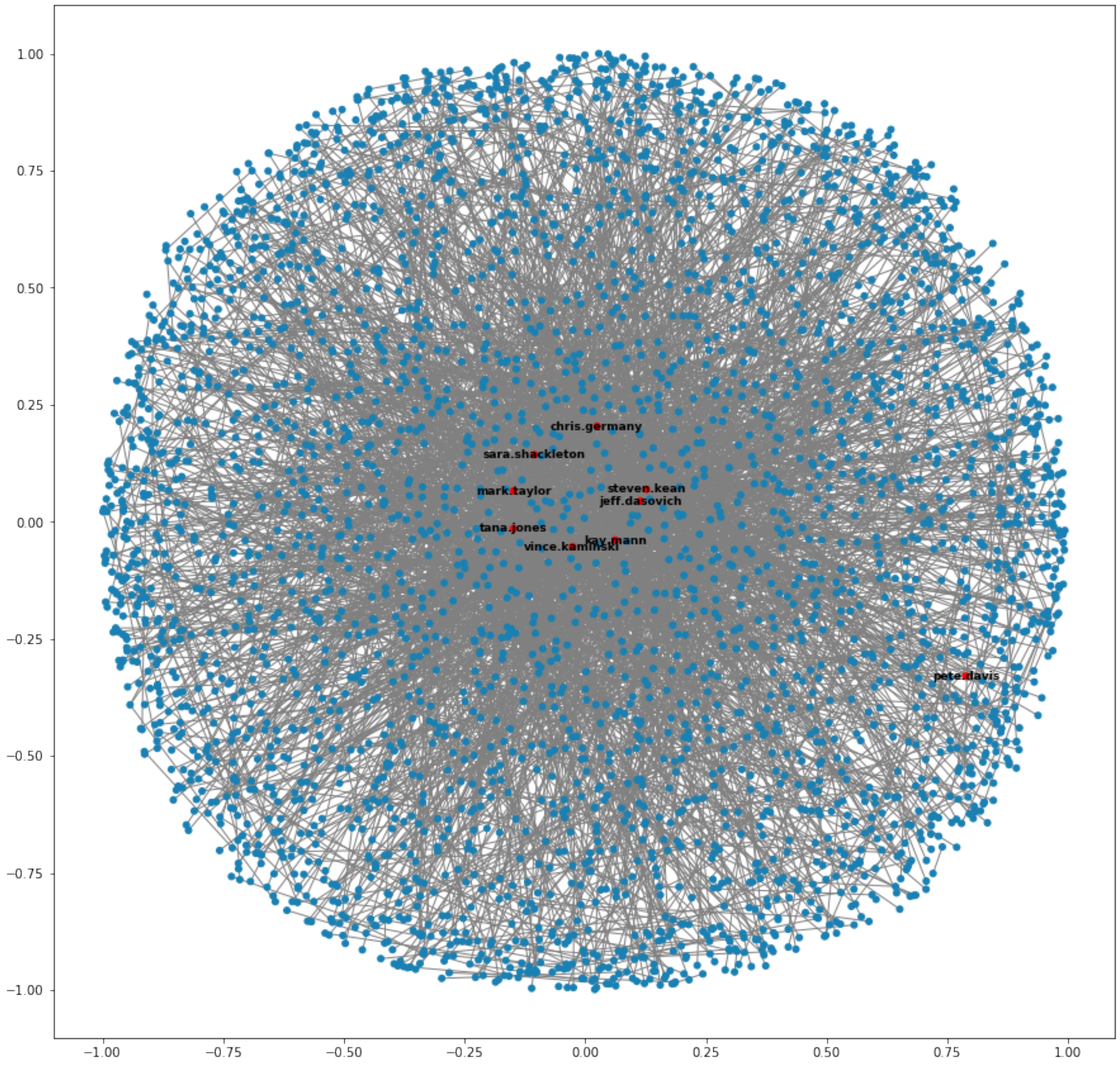


Figure 11: The connection between employees at ENRON

Here we highlighted the nodes representing the top senders in red. So that we can clearly see their position in the graph.

4.2 How close are the employees to each other

The degree centrality of each node of the graph gives us an idea on how well each person is connected to the others.

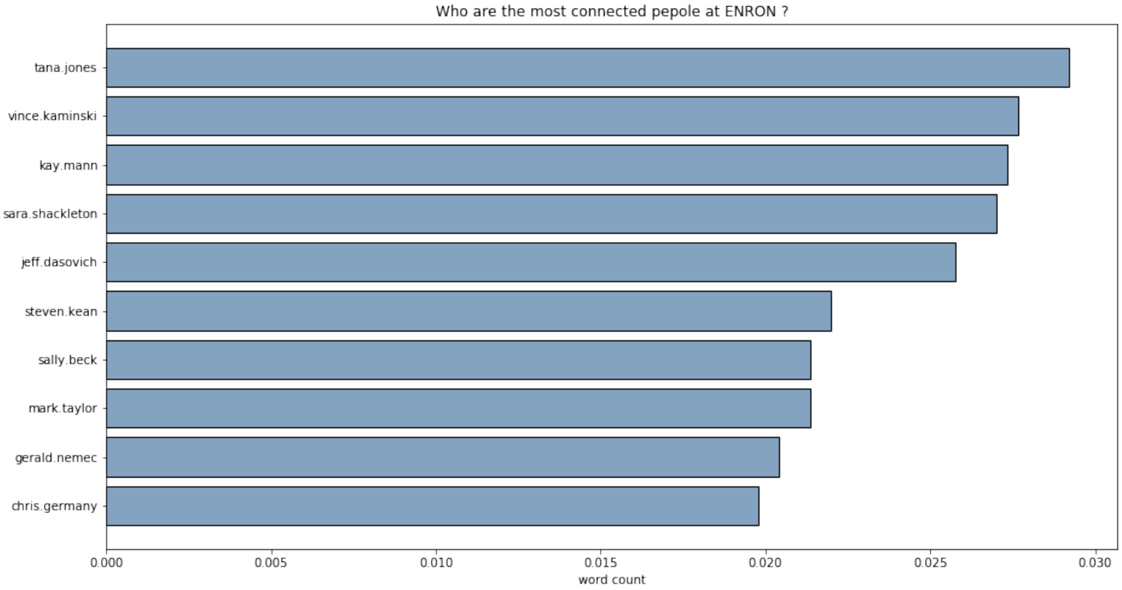


Figure 12: Degree centrality for the top employees

We can see here that Kay Mann, Tana Jones , Vince Kaminski and Jeff Dasovich were the most connected people inside ENRON, they are the people who have the maximum number of contacts within the company. This can let us think that these people are mostly responsible of the fraud but we will see later that it is not the case.

Their high connectivity to employees could be due to the fact that these people work in the human resources department and therefore are in the position where they have to send emails all the time.

5 Fraud Detection

5.1 Persons of interest

Aiming at working on a supervised model, we needed a particular target to predict. We therefore used an article[6] that appeared on usatoday that was mentioning the names of all employees who either have a settlment with the government, pleaded guilty or testified in exchange of persecution immunity.
Those people have been proved to be in relationship with the company's fraudulent activities so we labeled them as People of interest and added a target column POI to our dataframe.

We will apply machine learning techniques using the e-mails dataset above and the financial dataset to build a model capable of detecting the people that were accused and why not reveal new corrupted employees that escaped from justice.

5.2 Data preprocessing

We tried to find matches between employees in the financial data-set and the emails dataset. Sadly, our inference of the e-mails of Enron employees from their name and surname was not always right, we were therefore obliged to look for some emails patterns between the e-mails of both datasets to match them. Despite our efforts, we see that many employees are still not matched. We were therefore obliged to look for a match between the two dataframes and add them manually.

After some manipulations, we found that some e-mails did not match simply because the employee had two names or simply because some employees chose to have their nicknames in their e-mail address like Rick for Richard or Josh for John.
Despite our investigation efforts to find e-mail adresses that match, we left a small number (around 20 employees) untraced.

To improve our financial dataset with the content of e-mails, we thought that it was a good idea to take into account the number of emails sent and received by employees which is a good indicator of the importance of an employee at Enron.
We also thought that it was a good idea to look at the communication with the POIs, we therefore added two features the number of emails sent to POIs and the number of emails received from POIs.

5.3 Feature Engineering

Aiming at putting more weight on e-mails contents, we assumed that the high presence of words denoting fear such as ’trial’, ’fraud’ or ’investigation’ in emails could indicate well if a person was of interest or not. We therefore added a feature called ’stress-score’ counting the number of words denoting fear used by employees in their communications. As we can see in the figure 12, this score started to rise on July 2000, while the stock started to decrease only on January 2001. This count could gave an insight about the fraud 6 months before the public started to suspect a strange behaviour from the management.

Besides, we noticed that the features Loan Advances and Director fees contained mostly 0 meaning that they will not add much information to our models. We therefore preferred to drop them.

The final features will therefore be :

Salary : Reflects items such as base salary, executive cash allowances, and benefits payments

Bonus : Reflects annual cash incentives paid based upon company performance. Also may include other retention payments

Long Term Incentive : Reflects long-term incentive cash payments from various long-term incentive programs designed to tie executive compensation to long-term success as measured against key performance drivers and business objectives over a multi-year period, generally 3 to 5 years

Deferred Income : Reflects voluntary executive deferrals of salary, annual cash incentives, and long-term cash incentives as well as cash fees deferred by non-employee directorsunder a deferred compensation arrangement

Deferral Payments : Reflects distributions from a deferred compensation arrangement due to termination of employment or due to in-service withdrawals as per plan provisions

Expenses : Reflects reimbursements of business expenses. May include fees paid for consulting services

Expenses : Reflects reimbursements of business expenses. May include fees paid for consulting services

Exercised Stock options : Reflects amounts from exercised stock options which equal the market value in excess of the exercise price on the date the options were exercised either through cashless (same-day sale), stock swap or cash exercises

Restricted Stock : Reflects the gross fair market value of shares and accrued dividends

Nb-Emails-Sent : The total number of e-mails sent

Nb-Emails-Received : The total number of e-mails received

Nb-Emails-Sent-FromPOI: The total number of e-mails received from POI

Nb-Emails-Sent-ToPOI: The total number of e-mails sent to POIs

Nb-Emails-Sent-FromPOI: The number of stress words in sent e-mails

Stress-Score: The number of stress words in sent e-mails

In addition we remarked that there were a lot of zeros in the number of emails sent and received for some employees due to the fact that we were unable to find e-mail addresses of some of the employees of the financial data. By inspecting their data, we noticed that all of them don't get salaries from Enron and only have some stock options. We assumed that these people are shareholders and not Enron employees and therefore dropped them.

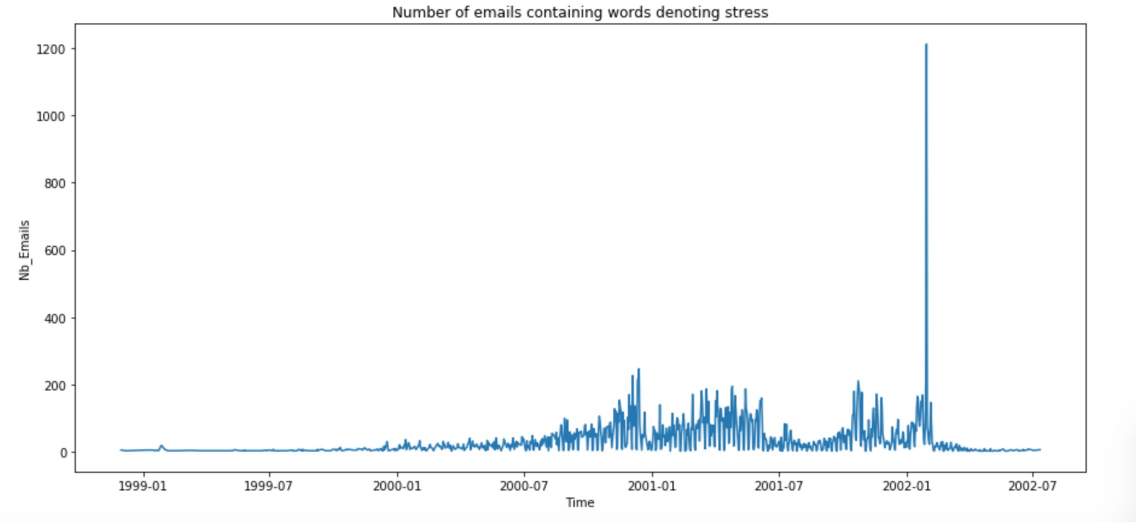
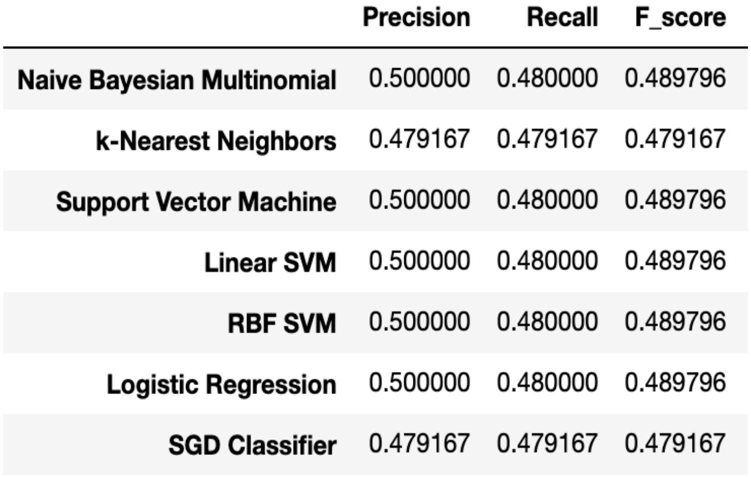


Figure 13: Stress score evolution in time

5.4 Supervised Learning Models

After cleaning our data set and enriching it with new features, we tried different machine learning models. All of them gave very high accuracies but since our target value was binary and unbalanced, accuracy was not the right measure of success. We therefore considered precision and recall as our main measures which appeared to be very poor for all models as we can see in Figure 13. This was due to the very unbalanced data we had at hand since only 18 employees were labeled POIs over 122 employees representing only 14 percent. We therefore explored two different techniques to overcome this issue.



	Precision	Recall	F_score
Naive Bayesian Multinomial	0.500000	0.480000	0.489796
k-Nearest Neighbors	0.479167	0.479167	0.479167
Support Vector Machine	0.500000	0.480000	0.489796
Linear SVM	0.500000	0.480000	0.489796
RBF SVM	0.500000	0.480000	0.489796
Logistic Regression	0.500000	0.480000	0.489796
SGD Classifier	0.479167	0.479167	0.479167

Figure 14: Prediction results

5.4.1 1st Idea : Over-Sampling

Over-Sampling increases the number of instances in the minority class by replicating them in order to present a higher representation of the minority class in the sample. We duplicated 3 times the rows containing POIs in our training set which led to a new data set of 40 percent of POIs. This was not the best idea since results got poorer for the different classifiers except for K Nearest Neighbors which reached 82 percent in both recall and precision. Figure 14 shows a confusion matrix of the results over a small test size of 35 samples.

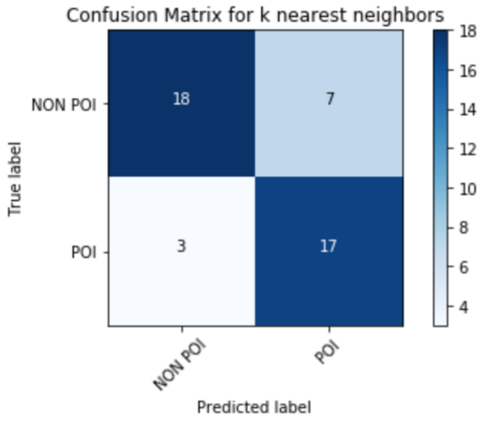


Figure 15: KNN results after over sampling

5.4.2 2nd Idea : Decision Trees

The second idea was to use machine learning models that were not affected by unbalanced data basically decision trees and random forests. Thanks to these models we were able to reach a maximum precision of 83 percent and maximum recall of 95 percent despite the high unbalancy. Figure 15 shows the results of applying this model on a test set of 25 samples.

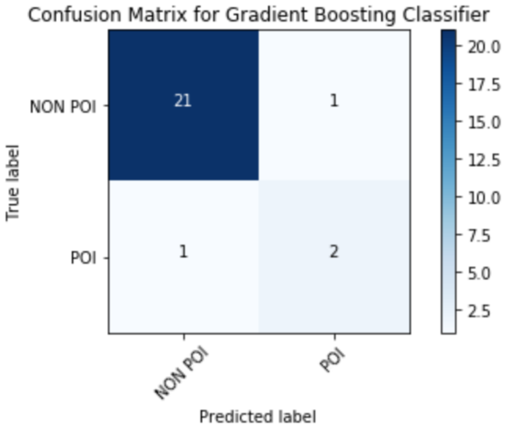


Figure 16: Gradient boosting classifier results

5.5 Unsupervised Learning Models

Trying to add more value to our solution, we also wanted to build a machine learning model capable of predicting fraudulent employees without any prior knowledge. We therefore used K-MEANS algorithm trying to cluster employees in two different groups, the fraudulents and the non fraudulents. With very low amount of data and without any prior knowledge of the scandal, we didn't put high hopes on the model. Our intuition was not right since the model was able to predict the two main responsible of the fraud : Lay Kenneth : founder, CEO and Chairman of Enron Corporation for most of its existence and Jeffrey Skilling : CEO of Enron Corporation during the scandal.

6 Conclusions

The Enron scandal gave us an opportunity to conduct studies on a real life email data-set that would have been very hard to come by because of a lot of regulations and privacy reasons.

From data exploration to network analysis and Fraud detection, we discussed what kind of analysis one can do on this type of data. We have had hands on experience on how to collect data from every source imaginable and how hard it could be to find clean data related to a specific subject.

We experimented with a lot of machine learning algorithms and learned how to correctly choose a model that outperforms every other one. We never thought we could achieve such impressive results. Our aim from the beginning was to build a tool for regulators to help them discover new scandals or develop a way for the company to find out if there is something shady going on under its roof(In this case it was the CEO who was corrupted) We are deeply confident that in the future the work of legal departments and regulator institutions will need more and more expertise in Data Analysis in order to reach new levels in the war against financial criminals.

References

[1] *The Enron Scandal*,
https://en.wikipedia.org/wiki/Enron_scandal

[2] *Enron Corpus*,
https://en.wikipedia.org/wiki/Enron_Corpus

[3] *Emails dataset*,
https://www.cs.cmu.edu/~enron/

[4] *Findlaw : Legal information.*,
https://www.findlaw.com/

[5] *Productivity Research*,
https://eprints.lse.ac.uk/4963/1/daysoftheweek\%28LSEROversion\%29.pdf

[6] *financial data*,
https://github.com/MayukhSobo/EnronFraud/blob/master/data/final_project_dataset.pkl

[7] *People of interest article*,
http://usatoday30.usatoday.com/money/industries/energy/2005-12-28-enron-participant
x.htm

[8] *pipedrive*,
https://www.pipedrive.com/en/blog/sales-productivity-around-the-world

[9] *California Electricity crisis*,
https://en.wikipedia.org/wiki/California_electricity_crisis