Identify Fraud from Enron Email

1. Introduction

Established in 1985, Enron became one of the largest companies in the world in 2000. In December 2001 it filed for bankruptcy, obliterating thousands of jobs and $60 billion in market value.
Investigation and trial lasted four years.

Arthur Andersen audit company was involved in the scandal. In 2002, the firm voluntarily surrendered its licenses to practice as Certified Public Accountants in the United States after it was found guilty of crimes in the firm's auditing of Enron.

Enron scandal was a major milestone for modern ethics and compliance rules :
In reaction to this major corporate and accounting scandal together with others in the same period, as Worldcom, a law was voted, called Sarbanes–Oxley Act , also known as the "Public Company Accounting Reform and Investor Protection Act". The sections of the bill cover responsibilities of a public corporation's board of directors, add criminal penalties for

certain misconduct, and require the Securities and Exchange Commission to create regulations to define how public corporations are to comply with the law.

The investigation into Enron's collapse was conducted by the Enron Task Force, a team of federal prosecutors within the Justice Department's Criminal Division, and agents from the FBI and the Internal Revenue Service Criminal Investigations Division. The Enron Task Force also coordinated with the Securities and Exchange Commission. The Enron Task Force was part of the President's Corporate Fraud Task Force, created in July 2002 to investigate allegations of fraud and corruption at U.S. corporations.

Enron Task force made significant amount of information public record, including tens of thousands of emails and detailed financial data for Enron employees. Dataset is made of this information.
In addition to Enron dataset, a list of Persons Of Interest (POI) was created manually. Is a POI either :someone who was indicted, or who settled without admitting guilt or who testified in exchange for immunity.

Goal of this project is to build a person of interest identifier based on financial and email data made public as a result of the Enron scandal. For this, as the amount of data available is hudge, machine learning tools are a must.

2. Dataset

Aim of this project is to define features that allow to identify person of interest.
This is meant to identify features that could help to identify future fraudsters, dishonnest employees, managers or consultants.

2.1Dataset source and python modules used.

In order to be in line with udacity « machine learning » explanations, we will use python 2.7 version together with skitlearn version 0.18 :

However, as both those versions are now decommissioned all project was run in python 3 and scikit learn 0.24. Only file poi-id.py can be run in python 2.7.

We will use Enron email dataset, in its 2015 version downloaded from link below :
https://www.cs.cmu.edu/~./enron/enron_mail_20150507.tar.gz
Email and finance data are combined into a single dataset, that we will explore in in this project.

2.2 Enron dataset : Persons Of Interest (Poi)

Among 146 records in the dataset, 18 relate to persons of interest : ('poi')

- Enron had 29000 employees in 2002, so 146 records of which 18 of POI class could seem not a large sample. However, 18 persons of interest seems a significant proportion in the sample : 12%. In addition, each record correspond to a person and gathers a lot of emails for each.

- For needs of machine learnings models, we will have to be careful when choosing training sample and test samples, so that to select significant samples : 12% of POI mean that we have to consider same % in the training set so that to have significant results.
- Considering the size of present sample, any conclusion would have to be validated with a larger sample.

About POI in our dataset :
Here is information about their former function in Enron and how they were sentenced.
We can see that all POIs of the dataset seem not to have the same level in implication in fraud.

POI's name	Function in Enron	Sentence
BELDEN TIMOTHY N	Head of trading in Enron Energy Services	2 years in prison
BOWEN JR RAYMOND M	Treasurer and chief financial officer of the Enron Corp	settlement (500 000 USD)
CALGER CHRISTOPHER F	Enron vice president	guilty plea withdrawn based on Honest Services.
CAUSEY RICHARD A	Chief accounting officer (former head of Enron audit team in Arthur Andersen's)	released
COLWELL WESLEY	Chief accounting officer of Enron Wholesale Service (former Arthur Andersen's auditor)	settlement (500 000 USD)
DELAINEY DAVID W	Head of Enron North America,	30 months in prison
FASTOW ANDREW S	Chief financial officer	10 years in prison
GLISAN JR BEN F	Corporate treasurer (former Arthur Andersen's employee)	5 years in prison
HANNON KEVIN P	Chief operating officer of Enron Corp.'s broadband Internet division	2 years in prison
HIRKO JOSEPH	co-chief executive officer of Enron Broadband Services	16 months in prison
KOENIG MARK E	Director of investor relations	10 years in prison
KOPPER MICHAEL J	Managing director in Fastow's finance division	37 months
LAY KENNETH L	Founder, CEO and Chairman	died before being sentenced
RICE KENNETH D	Chief executive of its high-speed Internet unit	27 months
RIEKER PAULA H	Corporate secretary	2 years probation
SHELBY REX	Senior vice president of engineering operations at Enron Broadband Services	2 years probation
SKILLING JEFFREY K	Chief executive officer	24 years in prison changed to 14 years

2.3 Enron dataset : features and labels

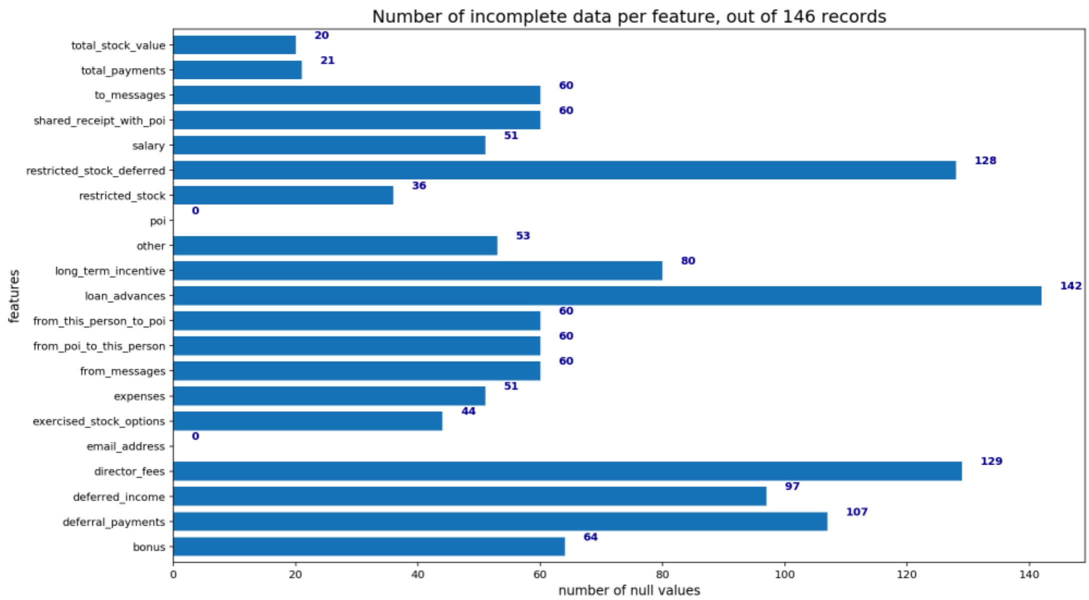
The dataset, contains 146 records .consisting of 20 features and 'poi'label
There are 14 features related to monetary information (« MONEY ») and 6 about mail content (« MAIL »)
Features_list : <class 'pandas.core.frame.DataFrame'>
Index: 146 entries, ALLEN PHILLIP K to YEAP SOON

Data columns (total 21 columns):	My feature/label classification
bonus	MONEY
deferral_payments	MONEY
deferred_income	MONEY
director_fees	MONEY
email_address	MAIL
exercised_stock_options	MONEY
expenses	MONEY
from_messages	MAIL
from_poi_to_this_person	MAIL
from_this_person_to_poi	MAIL
loan_advances	MONEY
long_term_incentive	MONEY
other	MONEY

poi	LABEL
restricted_stock	MONEY
restricted_stock_deferred	MONEY
salary	MONEY
shared_receipt_with_poi	MAIL
to_messages	MAIL
total_payments	MONEY
total_stock_value	MONEY

2.4 Dataset : removal of non significant records

2.4.1 NAN records



From dataset documentation, we read that values of NaN represent 0 and not unknown quantities.

We can see that three features have a lot of null values :

- restricted_stock_deffered ;
- loan advances ;
- directors fees ;

In particular, let's have a look at non zero loan advances records :

	loan_advances
FREVERT MARK A	2000000
LAY KENNETH L	81525000
PICKERING MARK R	400000

Number of non zero values shows that Mark Frevert and Mark Pickering, Enron executives who were not involved in Enron scandal,received loans that were high but much lower than that of Kenneth Lay (81M USD !!).
The fact that only three records are significant allows us not to choose « loan_advances » as a feature in our test. This feature seems not to be very significant for our analysis : This can not help to identify other PoI.

Other two features are kept : also there are not a large amount of information, those two information might be of some interest.

We will then remove records with no significant data :
Let's identify records where more than 19 features are NAN and remove them from the dataset.
After removal, dataset contains 141 records and 20 features.

From this point of the analysis, all NaN values in money features are replaced by 0.

2.4.2 Low variance features

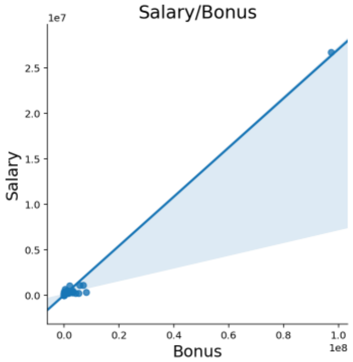
We will also eliminate features with low variances.
We remove from dataset all records where less than 19 data is filled in.
GRAMM WENDY L
LOCKHART EUGENE E
THE TRAVEL AGENCY IN THE PARK
WHALEY DAVID A,
WROBEL BRUCE

2.4.3 Errors

When looking at some features content, we could see that there were values that were « strange » : negative payments or total_payments feature that seemed different from payment arithmetical sum. We identify two wrong entries from the official pdf document and correct them.

	BELFER ROBERT	BHATNAGAR SANJAY
salary	0	0
to_messages	0	523
deferral_payments	-102500	0
total_payments	102500	0

2.5 Outliers identification and removal



A scattered plot together with list of bonus and salaries per person reveals outlier to be obviously removed from dataset : Total

bigest salaries, bonus, with POI indication

	bonus	salary	poi
TOTAL	97343619	26704229	0
LAVORATO JOHN J	8000000	339288	0
LAY KENNETH L	7000000	1072321	1
SKILLING JEFFREY K	5600000	1111258	1
BELDEN TIMOTHY N	5249999	213999	1
ALLEN PHILLIP K	4175000	201955	0
KITCHEN LOUISE	3100000	271442	0
DELAINEY DAVID W	3000000	365163	1
MCMAHON JEFFREY	2600000	370448	0
FALLON JAMES B	2500000	304588	0
FREVERT MARK A	2000000	1060932	0
SHANKMAN JEFFREY A	2000000	304110	0
RICE KENNETH D	1750000	420636	1
HICKERSON GARY J	1700000	211788	0
SHERRIFF JOHN R	1500000	428780	0
HANNON KEVIN P	1500000	243293	1
BOWEN JR RAYMOND M	1350000	278601	1
FASTOW ANDREW S	1300000	440698	1
CALGER CHRISTOPHER F	1250000	240189	1
COLWELL WESLEY	1200000	288542	1
BAXTER JOHN C	1200000	267102	0
HAEDICKE MARK E	1150000	374125	0
MCCONNELL MICHAEL S	1100000	365038	0
MULLER MARK S	1100000	251654	0

After removal, remaining outliers values appear to be that of most severly sentenced POI.
But.. Among some of the top ranking bonus, we can find non POI.
Internet search about John Lavorato leads to an article in the British newspaper « the Guardian », that shows why John Lavorato and Louise Kitchen received such bonuses and were not indicted :

Quote :« The doubts were raised about Mr Skilling's testimony on the same day as CNN revealed that some 500 Enron staff had received windfalls ranging from $1,000 to $5m. The payments were made to retain staff as the firm faced collapse. To get the cash, the staff agreed to stay for 90 days.

The highest payment of $5m went to John Lavorato, who ran Enron's energy trading business, while Louise Kitchen, the division's British-born chief operating officer, pocketed $2m. Both have taken up new jobs with UBS Warburg, the investment bank that now owns the division. » (…)

Workers laid off by Enron have, by contrast, been paid the minimum severance of $4,500 before tax and many are struggling to find work. » end of Quote

We can see that both John Lavorato and Louise Kitchen ratio bonus/salary is much higher than that of Mark Frevert who was chief executive of Enron Europe from 1996 until 2000 and then appointed chairman of Enron in 2001. All three are not POI, according to Katie Malone manual list.
Are we trying to find a true identifier for not ethic members of Enron ? Or just trying to find an identifier for what lead Katie Malone to create the list ?
We will here, consider that we are looking for an indicator of what lead Katie to select people as Poi.

So, to remove « non pure Poi » is a way to make a more powerful model : we will remove John Lavorato and Louise Kitchen from the dataset.

This ratio bonus/salary, seems however to us an interesting feature to create. As those tremendous bonuses were granted a few months before bankrupt, this ratio marks desperate attempt to hide desastrous situation of the company. This sort of indicator could be an alert for SEC in future comparable situations.

Another interesting key indicator is the mail flow from and to POI. John Lavorato and Louise Kitchen were granted bonuses by well informed people about fraud, so, they were in contact with POI . So, **mail received and sent to POI in % of all mails** can be interesting too.

3. Features and classifiers – first exploration

Exploration of dataset enabled to remove outliers, gave hints for new features to create and others featrures to supppress.

3.1 Additional features creation

We create three additional features :
- % of mails from POI
- % of mails to POI
- salary/bonus ratio

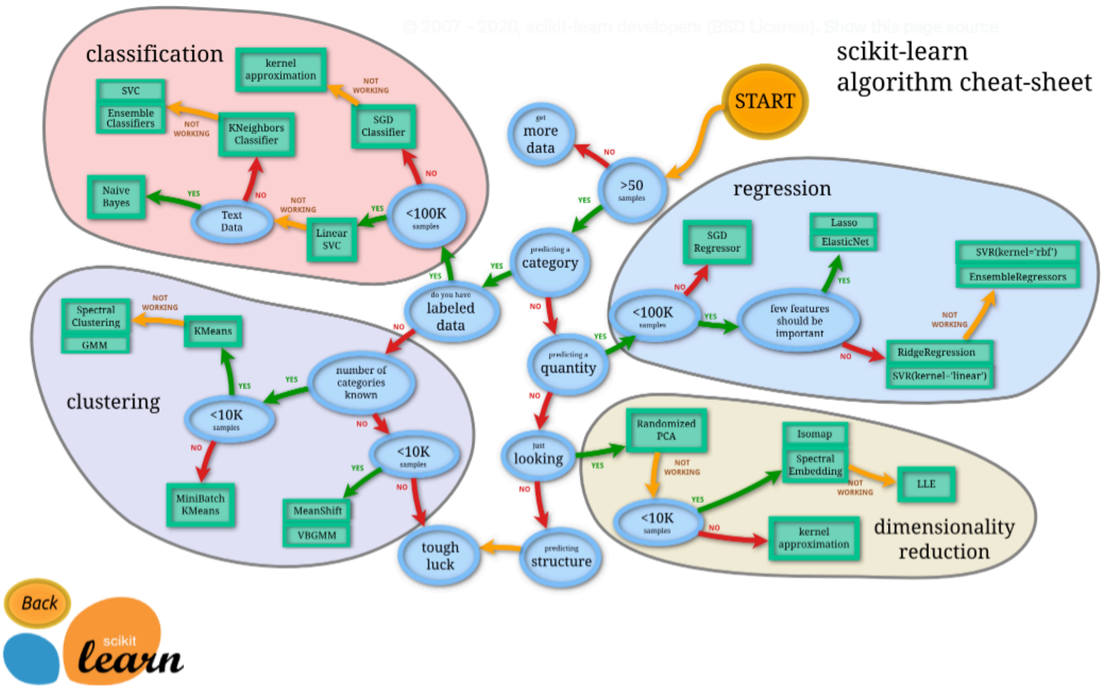
3.2 First exploration of the cleaned dataset, with DecisionTreeClassifier

We first create a trainset and test set with « train_test_split » from skitlearn library, with all features in the dataset.
number of POIs in trainset is 12.0 out of 98 that is in % : 12%
number of POIs in testset is 6.0 out of 43 that is in % : 14%

3.2.1 DecisionTreeClassifier choice

As DecisionTreeClassifier is a classifier that I understand well. It seems adapted to a small number of data, not being text, according to skitlearn documentation :

Source : https://scikit-learn.org/stable/tutorial/machine_learning_map/index.html



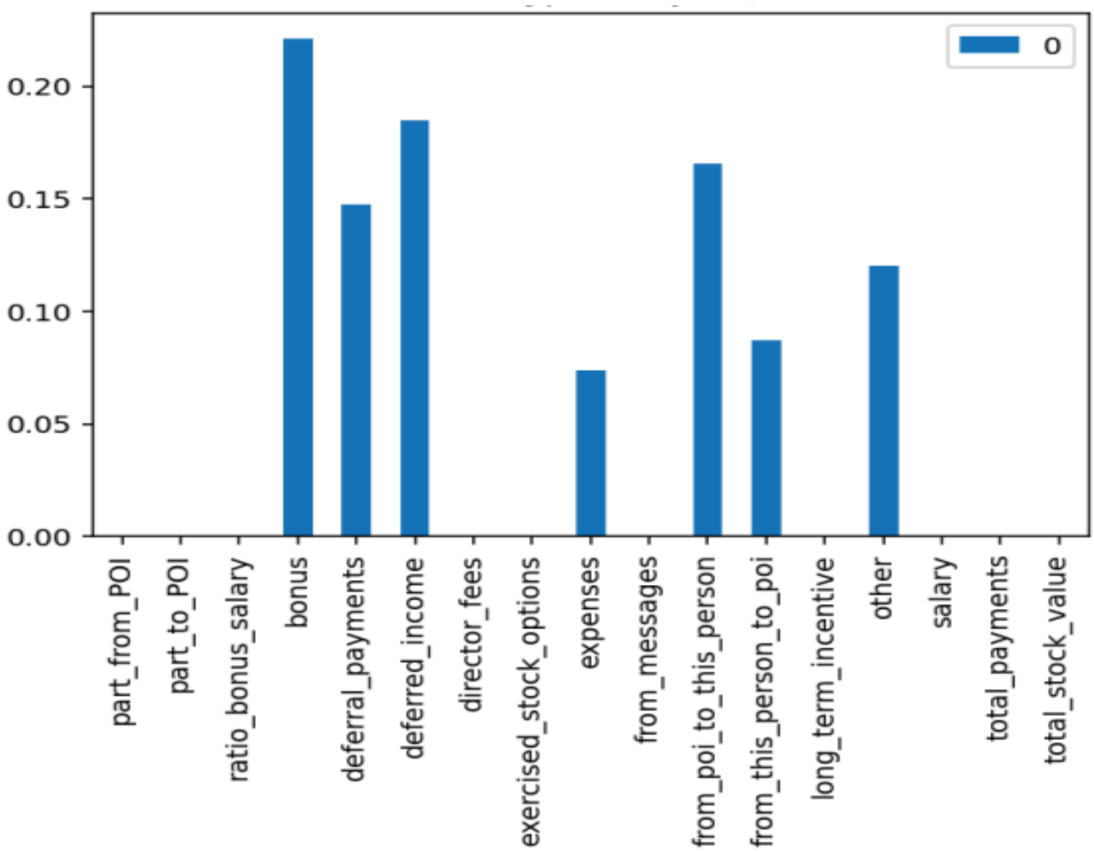
3.2.2 Features choice and importance

In order, not to neglect any feature that could be important, we will use for this first exploration, all features

I would expect one feature at least in every category of feature (mail activity, stock value and payments) to appear in most important features.

I would expect my three new features to appear in the most important features :

As regards features in each category, my intuition was good.
However, my new engineered features do not seem to work so well :
None of them is in the most important features :



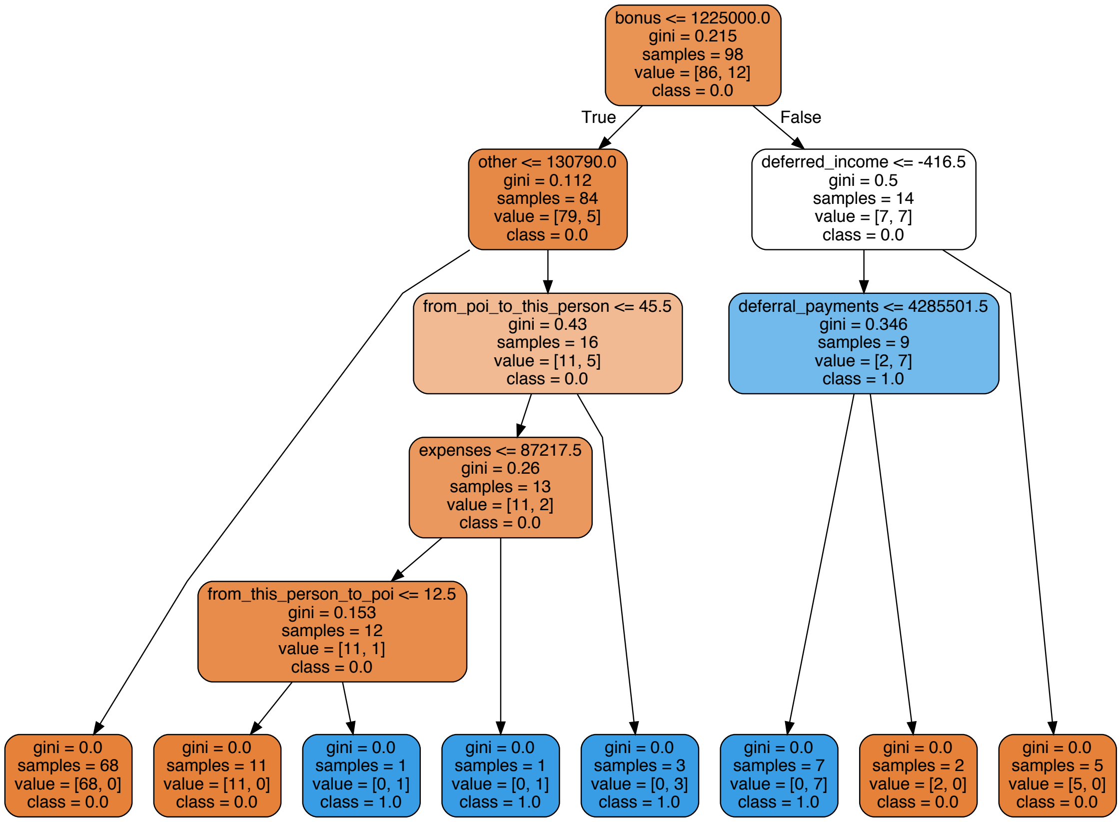
3.2.3 Decision tree graphical representation

Decision tree made with trainset shows where the decision nodes are.

I thought that what was important was ratio bonus/salary, because Enron employees were traders, the wages of whom are made of incentives. However, bonuses were so high for POI there is no use to refine here. 100% of chances of revealing a POI if he got a bonus greater than 1.6MUSD.
Scaling of features might there have an importance, especially when we will train trainset on SVC : there are a lot of outliers in the dataset.

. « Others » second feature in importance is made of various payment that can be used to increase personal profit without benefit to the company :« items such as payments for severence, consulting services, relocation costs, tax advances and allowances for employees on international assignment (i.e. housing allowances, cost of living allowances, payments under Enron's Tax Equalization Program, etc.). May also include payments provided with respect to employment agreements, as well as imputed income amounts for such things as use of corporate aircraft. »

. From POI to this person is the third most important feature : however it is not so important because, as others include consulting fees, it is understandable that consultants receive a lot of mails from POI , even if they are not POI.

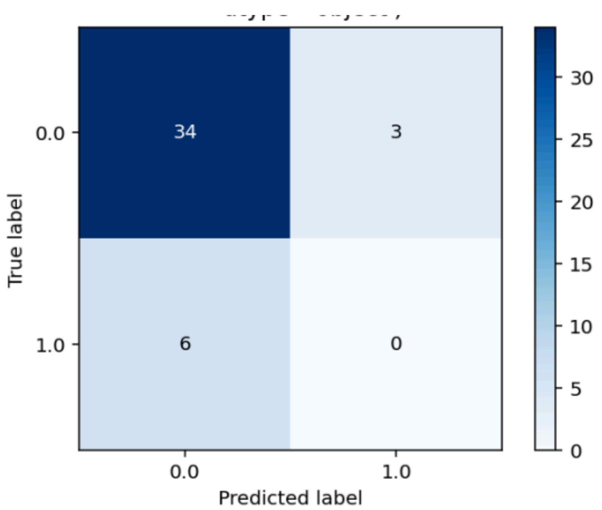


.

Results on test set

Let's see what are the result on test set :

Here is the confusion Matrix summarizing prediction made with model trained on trainset to test set :



POI recall is zero : no POI detected out of the 6 in the test set. Our objective is at least 30% on this criteria…

Non POI recall is better: 3 non POI identified as POI out of 36

Our model is putting in jail innocents but sets guilty people free. Not so good.

As a summary,

CLASSIFICATION REPORT

	precision	recall	f1-score	support
0.0	0.85	0.92	0.88	37
1.0	0.00	0.00	0.00	6
accuracy			0.79	43
macro avg	0.42	0.46	0.44	43
weighted avg	0.73	0.79	0.76	43

Accuracy, ie correct prediction over total population in macro average is 46%
Precision ie identification of positive prediction over total population to be predicted is 42% in average, 0% for POI, 85% for non POI
F1 score is the weighted average of precision and recall, so, in the same range as both previous scores.

Model has clearly to be improved because its scores are low in all dimensions…

Let's try to improve our model efficiency :

4. Selection of the best classifier, with most important features and best parameters

4.1 Classifiers

Sklearn documentation shows that we can improve our model by trying classifiers such as SVC models or KNeigborsclassifier or ensemble models.
Looking further to this documentation we will choose :
- Support vector model (linear being advised as first choice) => **SVM**
-KNeighborClassifier => **KNeighborClassifier**
- EnsembleClassifier averaging method : **RandomForestClassifier**
boosting method : **AdaBoostClassifier**

As I do not precisely understand all hyperparameters behind each model, I will iterate among parameters encountered during Udacity training and that I can understand.

4.2 Scaling

I can see that scaling is needed for support vector models
I guess that scaling could solve possible overfitting due to bonus feature.

4.3 Feature selection

I will try SelectKbest tool and iterate with k parameter

4.4 Pipeline and GridsearchCV

I create a pipeline with scaling, classifier, and SelectKbest parameters above.
I ran through GridsearchCV a cross validation using this pipeline.
Parameter for GridsearchCV are 3 crossvalidations and score to optimize : « f1» as we want to maximise recall and precision (more than 30%) and as f1 is weighted average of those two score.

4.5 Result

Pipeline(steps=[('scaler', MinMaxScaler()), ('selector', SelectKBest(k=7)), ('classifier', DecisionTreeClassifier())])

7 BEST_FEATURES with BEST MODEL
part_from_POI
part_to_POI
ratio_bonus_salary
director_fees
from_messages from_poi_to_this_person
total_payments
It seems that scaler has put aside bonus feature, still recal and precision are much better.

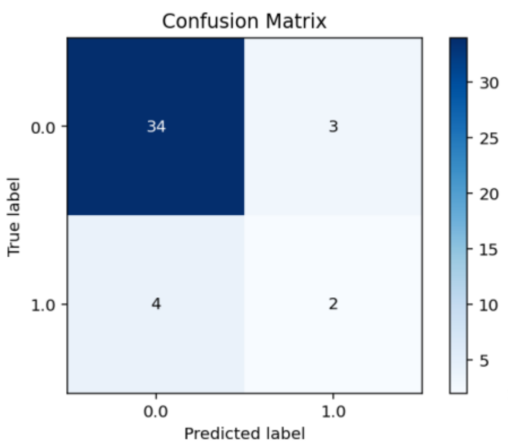
As I saw a lot of error messages, I am not sure however that this solution is the best one.

I tried to analyse results and understand what is going on, with cv_result dataframe, but I do not understand what really happens. It seems that there are a few samples out of crossvalidation that show a reasonnable percentage of POI. In the three selections of cross validation, result of recall is often Nan (division by 0)
I understand that chi2 for SelectKbest is also one of the causes for dumping, but I do not understand why.
I do not understand why an error message requesting scaler is there, as scalers are proposed.

This said, score of my model has significantly improved :

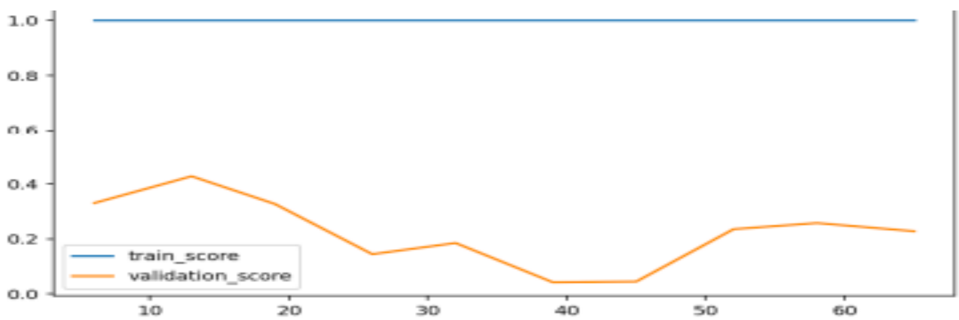
CLASSIFICATION REPORT

	precision	recall	f1-score	support
non POI	0.89	0.92	0.91	37
POI	0.40	0.33	0.36	6
accuracy			0.84	43
macro avg	0.65	0.63	0.64	43
weighted avg	0.83	0.84	0.83	43



I have now one third of POI identified but two false POIs are classified as POI.

Learning curve with recall perfo on trainset(in blue) and testset (orange) shows that model could be overfitted, also good results can be seen on tests. Perhaps problem is that samples chosen for trainset by cross validation often have very small POI number.



In a new similar project, I would not proceed in the same way as I did for this project : this time, I tried to guess best features and best engineered features and adapt to what I have created by intuition (and inflence of udacity training small projects). I think a better way to proceed is first to clean major errors, and transform dataset in numerical data, then to choose with sklearn map best adapted classifiers, run one of them to then find best features and think then about new features. Then run a pipeline combining, SelectKbest, feature scaling and seemingly most adapted classifiers

5. Conclusion

As a former auditor, in the 80s and now finance controller, I was very interested in this Enron story and this dataset.

During this project, with a very limited knowledge in data science, I could select a model that predicts reasonably well potential frauds in a company. This has been facilitated by the flexibility of the Python environment and the availability of powerful machine learning libraries.

In a few runs I got a model that seems to work with reasonable accuracy and precision. However this model is based in a limited dataset, and on a fuzzy definition of abnormal behaviors. A real application would require more data and a deeper analysis.
We would not fully track unethic or non compliant attitudes with the indicators of the current model. The model would have to be enriched with additional data and other features, such as keywords in mails or number of subsidiaries created and where key players are involved…

I am also sometimes unconfortable with the use of classifiers as a black box, without mastering the math inside (what is behind all parameters and behind some classifiers themselves, even with sklearn documentation). I always wonder if there is a more appropriate model. Moreover I wonder whether the classifiers have been properly validated.

Fortunately this can be counterbalanced with common sense and domain knowledge and a systematic critical review of the results.

.