

## Enron Submission Free-Response Questions

### 1. Project goal, machine learning utility, dataset background and use, outlier identification and handling

The purpose of this project is to choose an algorithm and fine tune it to get identification of Person of interest in Enron fraud. It is a supervised learning as we have the solution and we could check the error of the algorithm. Once we get a set of data, the levers that we have to get a good result are the features selection and definition, the classifier and the fine tune of the classifier.

The project result is the identification of who could be considered as person of interest in the investigation of Enron fraud. At the time of the fraud investigation some financial figures from employees were revealed as well as the emails in their accounts. The idea is to use the information related to financial compensations and emails exchange to define who is and who is not a person of interest for the investigation.

This could be considered as a prototype to enlarge the investigation over all the company employees to check if the investigation does not identify the people involved. We consider in this case the algorithm is impartial and designate the ones that should have been investigated.

The main limitation is that the designation of poi or non-poi is a result of not only the investigation but the pertinence of design someone as poi or non-poi, e.g. Some people involved could have passed an agreement with the judge to provide information in exchange of immunity, then will not be in the dataset as poi despite our algorithm identifying the person as poi.

Other limitation is that we have access to a limited amount of emails, not the complete corpus. The email information has been manually treated before, it means that we have access only to the ones that have been considered relevant for the investigation. We cannot be sure that cleaning has been properly done in the sense that some key message could have been deleted and other non-relevant remains in the set. On top, we have access only to the emails of Enron employees, the large number of entries (persons) have no emails in the dataset.

All that is real world, when we address any investigation or analysis we need to be prepared to deal with limited information with a certain level of non-quality. It is important to check the information we have and understand the information we do not have. We need to attack the analysis taking into account the limitations and the possible bias induced by that. Same applies to data quality.

We need always to keep in mind that the algorithm result could be modulated to get one or another result, depending on the objective of the analysis. The algorithm is impartial, but the information we pass to it modulates the result.

The key is the data quality and the data treatment before starting to submit the data set to the algorithm. We need to spend some time on data to understand what the matter is and to ensure the information we will pass to the algorithm will not induce bias, error or misleading results. We need to clean the data, identify the outliers and decide if we keep or remove them, define which features are useful and which ones we could create as combination of existing features. Finally we have to try different algorithms and fine tune them.

The information we have is a pkl file (final\_project\_dataset.pkl) that will be our initial data set and a pdf file with financial details per person (enron61702insiderpay.pdf) that will be useful to crosscheck information on the dataset and understand better the financial features and their possible relevance for the algorithm.

Exhibit B.1													
Insider	Payments							Stock Value (\$)					
	Salary (1)	Bonus (2)	Long Term Incentive (3)	Deferred Income (4)	Deferred Payments (5)	Loan Advances (6)	Other (7)	Exercises (8)	Unexercised Options (9)	Unexercised Restricted Stock (10)	Restricted Stock Deferred (11)	Total Stock Value (12)	
SHERRILL JOHNSON	438,760	1,500,000	754,432	-	-	-	1,871,138 (A)	4,335,158	1,831,578	1,209,454	-	11,336,828	
SHERRILL JOHNSON	631,028	800,000	8,000,000	-	-	0	50,000	0	6,862,780	10,000,000	6,849,870	26,694,650	
STABLE, BRYAN	280,760	750,000	-	-	-	-	270,071	-	1,162,514	-	-	2,112,344	
SULLIVAN-MARGALITZ, COLLIER	166,770	100,000	754,432	-	-	-	-	899,156	1,363,775	-	-	1,643,775	
SUNDER, MARTIN	377,440	700,000	400,000	-	-	-	-	1,343,059	-	690,020	-	688,020	
TAYLOR, MICHELLE S.	280,324	499,000	-	-	-	-	-	3,181,200	565,778	-	-	3,746,978	
TERRELL, TERENCE H.	121,000	-	279,433	-	-	16,784	-	91,433	4,412,476	350,120	-	4,813,796	
TERRELL, ALBERTA	273,030	-	273,030	-	(771,440)	-	-	390,835	-	576,822	-	1,468,657	
THOMPSON, ADAM S.	268,180	780,710	-	-	-	-	122,031 (B)	-	1,130,401	-	-	1,939,221	
TRUMPAUER, JONAS	-	-	-	-	(24,860)	-	-	538,122	228,018	-	-	766,140	
WARDMAN, JOHN	103,170	-	-	-	-	-	100,209	233,475	-	-	-	333,684	
WALLIS, R. ROBERT H.	371,041	830,000	940,771	-	-	-	2	10,806	4,245,544	1,572,433	-	5,868,897	
WATSON, LARRY W.	219,880	-	-	-	-	-	-	87,410	1,020,120	-	-	1,107,530	
WATSON, GEORGE	219,880	-	-	-	-	-	1,403	-	1,004,091	1,698,200	188,167	2,692,457	
WEINBAUM, EDWARD K.	61,244	-	276,191	(30,800)	-	-	-	763,151	96,718	194,890	-	264,890	
WHEATLEY, DAVID A.	100,000	-	-	-	-	-	-	100,000	-	-	-	200,000	
WHALLEY, TERENCE G.	87,394	5,000,000	838,134	-	-	-	-	4,977,174	3,282,060	3,768,177	-	6,079,137	
WILSON, T. WILFRED G.	670,000	630,000	-	-	-	-	-	670,000	6,000,000	6,000,000	-	12,670,000	
WILSON, N. HERBERT S.	-	-	-	-	(271,060)	-	-	1,413	108,579	84,092	-	193,671	
WOLFE, JOHN	-	-	-	-	-	-	-	108,161	-	108,161	-	216,322	
WOLFE, BRUCE	-	-	-	-	-	-	-	15,997	139,130	-	-	155,127	
WOLFE, S. SCOTT	670,000	0	0	0	0	0	0	670,000	6,000,010	6,000,000	-	12,670,010	
YEAR END	-	-	-	-	-	-	361,096 (C)	361,096	202,758	-	-	563,854	
TOTAL	516,796,780	187,243,610	3,450,510	517,891,000	517,885,000	553,813,000	\$61,497,120	\$1,101,100	\$3,209,431	\$313,764,000	\$230,123,000	\$17,474,780	\$442,609,331

The initial data set contains 146 entries (thereof 18 Poi and 128 non-poi) and 21 features, and the proportion of NaN goes from 14% (in features total stock value and total payment) to 97% (in feature loan advances). Regarding mails we have 24% missing emails address and 41% missing on exchange mails with poi.

The proportion of Nan for the 18 Poi is 100% for director fees and restricted stock deferred and 94% for loan\_advances

The proportion of Nan for the 128 non-Poi is 98% for loan\_advances, 87% for director fees and 86% for restricted stock deferred.

Thereof the non-financial features there are some interesting Nan:

- 100% of poi have email address
- 22% of poi have Nan for emails exchange despite they have email address
- 27% of non-poi do not have email address and 44% have Nan for emails

List of features, number of entries per feature (no-nan) and Nan proportion

	size	no-nan	nan-proportion
loan_advances	146	4	0.97
director_fees	146	17	0.88
restricted_stock_deferred	146	18	0.88
deferral_payments	146	39	0.73
deferred_income	146	49	0.66
long_term_incentive	146	66	0.55
bonus	146	82	0.44
from_messages	146	86	0.41
from_poi_to_this_person	146	86	0.41
from_this_person_to_poi	146	86	0.41
shared_receipt_with_poi	146	86	0.41
to_messages	146	86	0.41
other	146	93	0.36
salary	146	95	0.35
expenses	146	95	0.35
exercised_stock_options	146	102	0.30
restricted_stock	146	110	0.25
email_address	146	111	0.24
total_payments	146	125	0.14
total_stock_value	146	126	0.14
poi	146	146	0.00

Poi entries: List of features, number of entries per feature (no-nan) and Nan proportion

	size	no-nan	nan-proportion
director_fees	18	0	1.00
restricted_stock_deferred	18	0	1.00
loan_advances	18	1	0.94
deferral_payments	18	5	0.72
deferred_income	18	11	0.39
long_term_incentive	18	12	0.33
exercised_stock_options	18	12	0.33
from_this_person_to_poi	18	14	0.22
from_messages	18	14	0.22
from_poi_to_this_person	18	14	0.22
shared_receipt_with_poi	18	14	0.22
to_messages	18	14	0.22
bonus	18	16	0.11
salary	18	17	0.06
restricted_stock	18	17	0.06
poi	18	18	0.00
email_address	18	18	0.00
total_stock_value	18	18	0.00
other	18	18	0.00
expenses	18	18	0.00
total_payments	18	18	0.00

Non-Poi entries: List of features, number of entries per feature (no-nan) and Nan proportion

	size	no-nan	nan-proportion
loan_advances	128	3	0.98
director_fees	128	17	0.87
restricted_stock_deferred	128	18	0.86
deferral_payments	128	34	0.73
deferred_income	128	38	0.70
long_term_incentive	128	54	0.58
bonus	128	66	0.48
from_messages	128	72	0.44
from_poi_to_this_person	128	72	0.44
from_this_person_to_poi	128	72	0.44
shared_receipt_with_poi	128	72	0.44
to_messages	128	72	0.44
other	128	75	0.41
expenses	128	77	0.40
salary	128	78	0.39
exercised_stock_options	128	90	0.30
restricted_stock	128	93	0.27
email_address	128	93	0.27
total_payments	128	107	0.16
total_stock_value	128	108	0.16
poi	128	128	0.00

The interpretation of NaN is different depending on the features. Financial features NaN means that the person have got 0 on that feature, e.g. only 3 entries with value in loan advances features means that only 3 person got money through that concept. Those NaN will be considered as 0.

We could consider the financial feature with a big amount of Nan as outlier and think about to remove from the data set, e.g. loan\_advances

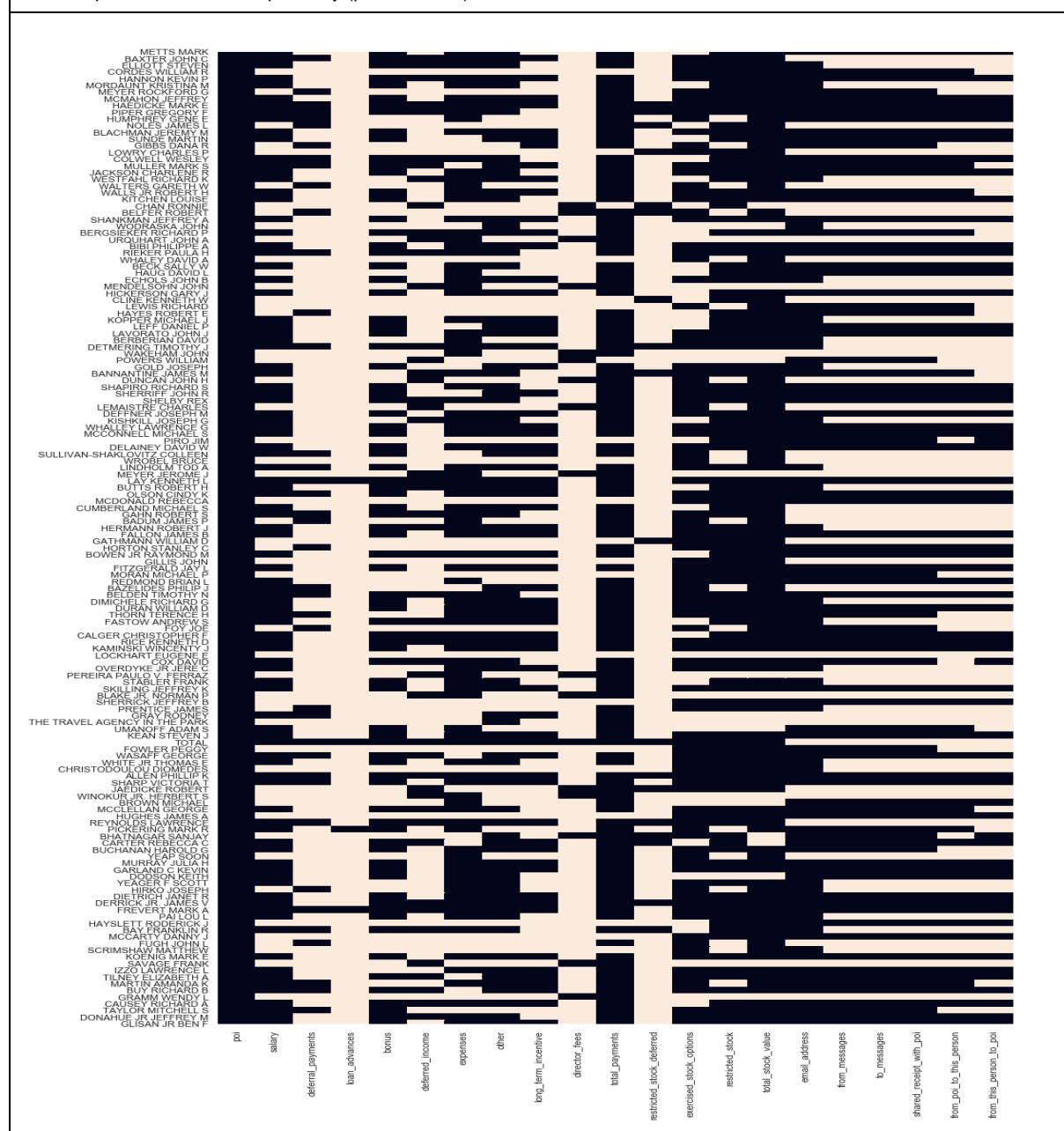
Regarding emails features NaN, it becomes more delicate to analyses, solve and make a decision.

We have 35 non-poi entries without emails address that means basically they do not work for Enron Company at that time but as they are in the pdf financial document, we see they have a kind of relation with the company. (Further details visible running PML\_1\_data\_adq.py)

There are different strategy to deal with Nan in data set. We could think about the option of imputing missing data but there will be difficult to define a figure to be accurate, average does not add meaning in this case as the emails amounts are not the real one but the result of a cleaning process. We will transform Nan in 0 and manage the entries depending on that.

A good way to have a global intuitive view of the level of Nan is represent them in a chart

Visual representation of NaN per entry (person name). In black the available information



This visualization give us some indications about the distribution of Nan per entry.

We should clean the dataset by analysis of Nan, identify the outliers and check the data quality

We run different analyses per features and entries, visible on PML\_1\_..., PML\_2\_..., PML\_3\_..., where you could read further details regarding the features and the entries if you want to enlarge your knowledge on the subject.

Hereby some of the findings:

- Entry with no Nan in finance features that is TOTAL. This is not a person itself but we could use to check the financial data and identify some mistakes in the financial data as that has been transfer manually from a pdf to the dataset and that could have generate mistakes. We will remove from the final data set
- 3 entries with wrong data, three for two entries ('BELFER ROBERT' and 'BHATNAGAR SANJAY') that have wrong values in payments and bonus, we could correct that as we have the pdf file with the right values, then we correct before continue the analysis
- Entry that is not person: 'THE TRAVEL AGENCY IN THE PARK' that we will remove from the final data set
- Entry "Eugene Lockhart" with all features Nan except False in poi. We could remove this person from the data set if we consider information added is 0 but the information behind is that if all values are 0, the entry is non-poi.

Hereby a table with the top 15 nan entries

	size	no-nan	nan-proportion
LOCKHART EUGENE E	21	1	0.95
GRAMM WENDY L	21	3	0.86
THE TRAVEL AGENCY IN THE PARK	21	3	0.86
WHALEY DAVID A	21	3	0.86
WROBEL BRUCE	21	3	0.86
WAKEHAM JOHN	21	4	0.81
GILLIS JOHN	21	4	0.81
CLINE KENNETH W	21	4	0.81
SAVAGE FRANK	21	4	0.81
SCRIMSHAW MATTHEW	21	4	0.81
WODRASKA JOHN	21	4	0.81
GATHMANN WILLIAM D	21	5	0.76
MEYER JEROME J	21	5	0.76
CHAN RONNIE	21	5	0.76
URQUHART JOHN A	21	5	0.76

All along the project some outlier in different features has been identified and investigation about them done (see PML\_1\_..., PML\_2\_... and PML\_3\_...).

LAY KENNETH L is an outlier for total payment and total stock value as he was the founder, CEO and Chairman of Enron at that time. We not to be removed from our data set.

'KAMINSKI WINCENTY J' is an outlier as 'from messages' and he is non-poi.

'SHAPIRO RICHARD S' and are outlier in 'to messages' and he is non-poi.

'KEAN STEVEN J' is second outlier for 'from messages' and 'to messages' and he is non-poi.

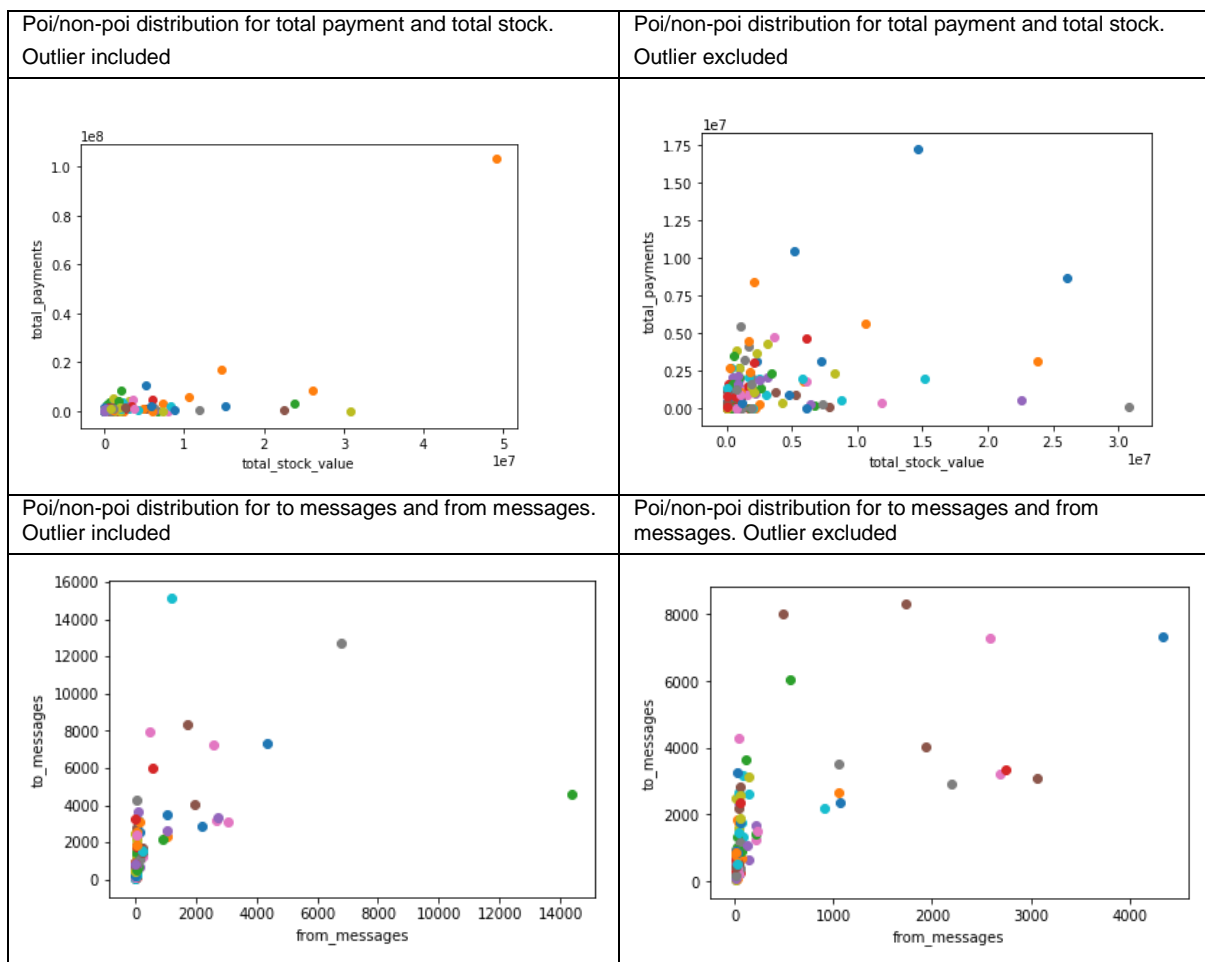
DELAINEY DAVID W is outlier with more than 600 mails 'from\_this\_person\_to\_poi' and he is poi.  
LAVORATO JOHN J is outlier with more than 500 mails 'from\_poi\_to\_this\_person' and he is non-poi.

We identify the amount of features where the top outlier in poi and non poi are outlier

POI outlier champions: amount of features		Non POI outlier champions: amount of features	
LAY KENNETH L	9	FREVERT MARK A	13
BELDEN TIMOTHY N	8	WHALLEY LAWRENCE G	10
SKILLING JEFFREY K	6	LAVORATO JOHN J	10
HIRKO JOSEPH	2	HAEDICKE MARK E	9
RICE KENNETH D	2	BAXTER JOHN C	8

We remove those outlier from the data set some times to run investigations but we do not find any indication that the values are mistake, then we do not remove form the data set as considered as relevant for the analysis.

The distribution of the entries and their relation changed depending if we keep the outlier or we remove them, that is visible in the next charts



We should take into account the existing of outliers when we decide which scaler we are going to use to scale the features.

We found that 4 poi have no mails information: FASTOW ANDREW S, KOPPER MICHAEL J, YEAGER F SCOTT and HIRKO JOSEPH. We could remove them from the data set as considered

as noise value but we do not have too much poi entries then we decide to kept them as valuable information in financial features Final data set could have 14 poi with 100% of information related to mails exchanges

The point is that we have very low number of poi and if we remove some of them the algorithm could struggle even more to get the right answer.

The final data set I will use contains 144 entries (thereof 18 Poi), and the proportion of NaN goes from 13% (total stock value) to 98% (loan advances). Regarding mails we have 23% missing emails address and 40% missing on exchange mails with poi.

List of features, number of entries per feature (no-nan) and Nan proportion	Poi entries: List of features, number of entries per feature (no-nan) and Nan proportion	Non-Poi entries: List of features, number of entries per feature (no-nan) and Nan proportion																																																																																																																																																																																																																																																																								
<table><thead><tr><th></th><th>size</th><th>no-nan</th><th>nan-proportion</th></tr></thead><tbody><tr><td>loan_advances</td><td>144</td><td>3</td><td>0.98</td></tr><tr><td>director_fees</td><td>144</td><td>15</td><td>0.90</td></tr><tr><td>restricted_stock_deferred</td><td>144</td><td>17</td><td>0.88</td></tr><tr><td>deferral_payments</td><td>144</td><td>37</td><td>0.74</td></tr><tr><td>deferred_income</td><td>144</td><td>49</td><td>0.66</td></tr><tr><td>long_term_incentive</td><td>144</td><td>65</td><td>0.55</td></tr><tr><td>bonus</td><td>144</td><td>81</td><td>0.44</td></tr><tr><td>from_messages</td><td>144</td><td>86</td><td>0.40</td></tr><tr><td>shared_receipt_with_poi</td><td>144</td><td>86</td><td>0.40</td></tr><tr><td>from_this_person_to_poi</td><td>144</td><td>86</td><td>0.40</td></tr><tr><td>to_messages</td><td>144</td><td>86</td><td>0.40</td></tr><tr><td>from_poi_to_this_person</td><td>144</td><td>86</td><td>0.40</td></tr><tr><td>other</td><td>144</td><td>90</td><td>0.38</td></tr><tr><td>salary</td><td>144</td><td>94</td><td>0.35</td></tr><tr><td>expenses</td><td>144</td><td>96</td><td>0.33</td></tr><tr><td>exercised_stock_options</td><td>144</td><td>100</td><td>0.31</td></tr><tr><td>restricted_stock</td><td>144</td><td>110</td><td>0.24</td></tr><tr><td>email_address</td><td>144</td><td>111</td><td>0.23</td></tr><tr><td>total_payments</td><td>144</td><td>123</td><td>0.15</td></tr><tr><td>total_stock_value</td><td>144</td><td>125</td><td>0.13</td></tr><tr><td>poi</td><td>144</td><td>144</td><td>0.00</td></tr></tbody></table>		size	no-nan	nan-proportion	loan_advances	144	3	0.98	director_fees	144	15	0.90	restricted_stock_deferred	144	17	0.88	deferral_payments	144	37	0.74	deferred_income	144	49	0.66	long_term_incentive	144	65	0.55	bonus	144	81	0.44	from_messages	144	86	0.40	shared_receipt_with_poi	144	86	0.40	from_this_person_to_poi	144	86	0.40	to_messages	144	86	0.40	from_poi_to_this_person	144	86	0.40	other	144	90	0.38	salary	144	94	0.35	expenses	144	96	0.33	exercised_stock_options	144	100	0.31	restricted_stock	144	110	0.24	email_address	144	111	0.23	total_payments	144	123	0.15	total_stock_value	144	125	0.13	poi	144	144	0.00	<table><thead><tr><th></th><th>size</th><th>no-nan</th><th>nan-proportion</th></tr></thead><tbody><tr><td>director_fees</td><td>18</td><td>0</td><td>1.00</td></tr><tr><td>restricted_stock_deferred</td><td>18</td><td>0</td><td>1.00</td></tr><tr><td>loan_advances</td><td>18</td><td>1</td><td>0.94</td></tr><tr><td>deferral_payments</td><td>18</td><td>5</td><td>0.72</td></tr><tr><td>deferred_income</td><td>18</td><td>11</td><td>0.39</td></tr><tr><td>long_term_incentive</td><td>18</td><td>12</td><td>0.33</td></tr><tr><td>exercised_stock_options</td><td>18</td><td>12</td><td>0.33</td></tr><tr><td>from_this_person_to_poi</td><td>18</td><td>14</td><td>0.22</td></tr><tr><td>from_messages</td><td>18</td><td>14</td><td>0.22</td></tr><tr><td>from_poi_to_this_person</td><td>18</td><td>14</td><td>0.22</td></tr><tr><td>shared_receipt_with_poi</td><td>18</td><td>14</td><td>0.22</td></tr><tr><td>to_messages</td><td>18</td><td>14</td><td>0.22</td></tr><tr><td>bonus</td><td>18</td><td>16</td><td>0.11</td></tr><tr><td>salary</td><td>18</td><td>17</td><td>0.06</td></tr><tr><td>restricted_stock</td><td>18</td><td>17</td><td>0.06</td></tr><tr><td>poi</td><td>18</td><td>18</td><td>0.00</td></tr><tr><td>email_address</td><td>18</td><td>18</td><td>0.00</td></tr><tr><td>total_stock_value</td><td>18</td><td>18</td><td>0.00</td></tr><tr><td>other</td><td>18</td><td>18</td><td>0.00</td></tr><tr><td>expenses</td><td>18</td><td>18</td><td>0.00</td></tr><tr><td>total_payments</td><td>18</td><td>18</td><td>0.00</td></tr></tbody></table>		size	no-nan	nan-proportion	director_fees	18	0	1.00	restricted_stock_deferred	18	0	1.00	loan_advances	18	1	0.94	deferral_payments	18	5	0.72	deferred_income	18	11	0.39	long_term_incentive	18	12	0.33	exercised_stock_options	18	12	0.33	from_this_person_to_poi	18	14	0.22	from_messages	18	14	0.22	from_poi_to_this_person	18	14	0.22	shared_receipt_with_poi	18	14	0.22	to_messages	18	14	0.22	bonus	18	16	0.11	salary	18	17	0.06	restricted_stock	18	17	0.06	poi	18	18	0.00	email_address	18	18	0.00	total_stock_value	18	18	0.00	other	18	18	0.00	expenses	18	18	0.00	total_payments	18	18	0.00	<table><thead><tr><th></th><th>size</th><th>no-nan</th><th>nan-proportion</th></tr></thead><tbody><tr><td>loan_advances</td><td>126</td><td>2</td><td>0.98</td></tr><tr><td>director_fees</td><td>126</td><td>15</td><td>0.88</td></tr><tr><td>restricted_stock_deferred</td><td>126</td><td>17</td><td>0.87</td></tr><tr><td>deferral_payments</td><td>126</td><td>32</td><td>0.75</td></tr><tr><td>deferred_income</td><td>126</td><td>38</td><td>0.70</td></tr><tr><td>long_term_incentive</td><td>126</td><td>53</td><td>0.58</td></tr><tr><td>bonus</td><td>126</td><td>65</td><td>0.48</td></tr><tr><td>from_messages</td><td>126</td><td>72</td><td>0.43</td></tr><tr><td>from_poi_to_this_person</td><td>126</td><td>72</td><td>0.43</td></tr><tr><td>from_this_person_to_poi</td><td>126</td><td>72</td><td>0.43</td></tr><tr><td>shared_receipt_with_poi</td><td>126</td><td>72</td><td>0.43</td></tr><tr><td>other</td><td>126</td><td>72</td><td>0.43</td></tr><tr><td>to_messages</td><td>126</td><td>72</td><td>0.43</td></tr><tr><td>salary</td><td>126</td><td>77</td><td>0.39</td></tr><tr><td>expenses</td><td>126</td><td>78</td><td>0.38</td></tr><tr><td>exercised_stock_options</td><td>126</td><td>88</td><td>0.30</td></tr><tr><td>restricted_stock</td><td>126</td><td>93</td><td>0.26</td></tr><tr><td>email_address</td><td>126</td><td>93</td><td>0.26</td></tr><tr><td>total_payments</td><td>126</td><td>105</td><td>0.17</td></tr><tr><td>total_stock_value</td><td>126</td><td>107</td><td>0.15</td></tr><tr><td>poi</td><td>126</td><td>126</td><td>0.00</td></tr></tbody></table>		size	no-nan	nan-proportion	loan_advances	126	2	0.98	director_fees	126	15	0.88	restricted_stock_deferred	126	17	0.87	deferral_payments	126	32	0.75	deferred_income	126	38	0.70	long_term_incentive	126	53	0.58	bonus	126	65	0.48	from_messages	126	72	0.43	from_poi_to_this_person	126	72	0.43	from_this_person_to_poi	126	72	0.43	shared_receipt_with_poi	126	72	0.43	other	126	72	0.43	to_messages	126	72	0.43	salary	126	77	0.39	expenses	126	78	0.38	exercised_stock_options	126	88	0.30	restricted_stock	126	93	0.26	email_address	126	93	0.26	total_payments	126	105	0.17	total_stock_value	126	107	0.15	poi	126	126	0.00
	size	no-nan	nan-proportion																																																																																																																																																																																																																																																																							
loan_advances	144	3	0.98																																																																																																																																																																																																																																																																							
director_fees	144	15	0.90																																																																																																																																																																																																																																																																							
restricted_stock_deferred	144	17	0.88																																																																																																																																																																																																																																																																							
deferral_payments	144	37	0.74																																																																																																																																																																																																																																																																							
deferred_income	144	49	0.66																																																																																																																																																																																																																																																																							
long_term_incentive	144	65	0.55																																																																																																																																																																																																																																																																							
bonus	144	81	0.44																																																																																																																																																																																																																																																																							
from_messages	144	86	0.40																																																																																																																																																																																																																																																																							
shared_receipt_with_poi	144	86	0.40																																																																																																																																																																																																																																																																							
from_this_person_to_poi	144	86	0.40																																																																																																																																																																																																																																																																							
to_messages	144	86	0.40																																																																																																																																																																																																																																																																							
from_poi_to_this_person	144	86	0.40																																																																																																																																																																																																																																																																							
other	144	90	0.38																																																																																																																																																																																																																																																																							
salary	144	94	0.35																																																																																																																																																																																																																																																																							
expenses	144	96	0.33																																																																																																																																																																																																																																																																							
exercised_stock_options	144	100	0.31																																																																																																																																																																																																																																																																							
restricted_stock	144	110	0.24																																																																																																																																																																																																																																																																							
email_address	144	111	0.23																																																																																																																																																																																																																																																																							
total_payments	144	123	0.15																																																																																																																																																																																																																																																																							
total_stock_value	144	125	0.13																																																																																																																																																																																																																																																																							
poi	144	144	0.00																																																																																																																																																																																																																																																																							
	size	no-nan	nan-proportion																																																																																																																																																																																																																																																																							
director_fees	18	0	1.00																																																																																																																																																																																																																																																																							
restricted_stock_deferred	18	0	1.00																																																																																																																																																																																																																																																																							
loan_advances	18	1	0.94																																																																																																																																																																																																																																																																							
deferral_payments	18	5	0.72																																																																																																																																																																																																																																																																							
deferred_income	18	11	0.39																																																																																																																																																																																																																																																																							
long_term_incentive	18	12	0.33																																																																																																																																																																																																																																																																							
exercised_stock_options	18	12	0.33																																																																																																																																																																																																																																																																							
from_this_person_to_poi	18	14	0.22																																																																																																																																																																																																																																																																							
from_messages	18	14	0.22																																																																																																																																																																																																																																																																							
from_poi_to_this_person	18	14	0.22																																																																																																																																																																																																																																																																							
shared_receipt_with_poi	18	14	0.22																																																																																																																																																																																																																																																																							
to_messages	18	14	0.22																																																																																																																																																																																																																																																																							
bonus	18	16	0.11																																																																																																																																																																																																																																																																							
salary	18	17	0.06																																																																																																																																																																																																																																																																							
restricted_stock	18	17	0.06																																																																																																																																																																																																																																																																							
poi	18	18	0.00																																																																																																																																																																																																																																																																							
email_address	18	18	0.00																																																																																																																																																																																																																																																																							
total_stock_value	18	18	0.00																																																																																																																																																																																																																																																																							
other	18	18	0.00																																																																																																																																																																																																																																																																							
expenses	18	18	0.00																																																																																																																																																																																																																																																																							
total_payments	18	18	0.00																																																																																																																																																																																																																																																																							
	size	no-nan	nan-proportion																																																																																																																																																																																																																																																																							
loan_advances	126	2	0.98																																																																																																																																																																																																																																																																							
director_fees	126	15	0.88																																																																																																																																																																																																																																																																							
restricted_stock_deferred	126	17	0.87																																																																																																																																																																																																																																																																							
deferral_payments	126	32	0.75																																																																																																																																																																																																																																																																							
deferred_income	126	38	0.70																																																																																																																																																																																																																																																																							
long_term_incentive	126	53	0.58																																																																																																																																																																																																																																																																							
bonus	126	65	0.48																																																																																																																																																																																																																																																																							
from_messages	126	72	0.43																																																																																																																																																																																																																																																																							
from_poi_to_this_person	126	72	0.43																																																																																																																																																																																																																																																																							
from_this_person_to_poi	126	72	0.43																																																																																																																																																																																																																																																																							
shared_receipt_with_poi	126	72	0.43																																																																																																																																																																																																																																																																							
other	126	72	0.43																																																																																																																																																																																																																																																																							
to_messages	126	72	0.43																																																																																																																																																																																																																																																																							
salary	126	77	0.39																																																																																																																																																																																																																																																																							
expenses	126	78	0.38																																																																																																																																																																																																																																																																							
exercised_stock_options	126	88	0.30																																																																																																																																																																																																																																																																							
restricted_stock	126	93	0.26																																																																																																																																																																																																																																																																							
email_address	126	93	0.26																																																																																																																																																																																																																																																																							
total_payments	126	105	0.17																																																																																																																																																																																																																																																																							
total_stock_value	126	107	0.15																																																																																																																																																																																																																																																																							
poi	126	126	0.00																																																																																																																																																																																																																																																																							

Some details on the entries with big amount of Nan could be seen in PML\_1 and PML\_2

The total payment average is 2.2m\$ with a Q3 on 1.9m\$ (lower than the average) with a maximum value of 103.56m\$. Total stock value has an average of 2.95m\$ and a Q3 at 2.31 \$ (lower than the average) with a maximum value of 49.1m\$. A Q3 lower than average means we have some outlier with a big delta versus the average value. That is thanks to our outlier K. Lay.

Regarding the amount of mails, we need to keep in mind that the emails has been cleaning but not deperate, that means there are mails not relevant for the investigation in the data set.

For further details and coding see PML\_1\_data\_adq.py and PML\_2\_data\_chc.py

2. Features: creation and explanation of rational for that new feature, selection process e.g. automated feature selection function like SelectKBest, scaling, feature importance as per algorithm used

The features in the data fall into two types namely financial features and email features, on top we have the POI labels that is the one we need the algorithm predict properly

POI label: ['poi'] (boolean, represented as integer)

Financial: ['salary', 'deferral\_payments', 'total\_payments', 'loan\_advances', 'bonus', 'restricted\_stock\_deferred', 'deferred\_income', 'total\_stock\_value', 'expenses', 'exercised\_stock\_options', 'other', 'long\_term\_incentive', 'restricted\_stock', 'director\_fees'] (all units are in US dollars)

Email: ['to\_messages', 'email\_address', 'from\_poi\_to\_this\_person', 'from\_messages', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi'] (units are number of emails messages; notable exception is 'email\_address', which is a text string)

After removing the people in explained in question 1 I made some analysis of the features, their relations and correlations, and the information we have for the features, I take some assumptions to remove features

The level on Nan is high for some features.

It seems logic to remove all the features with more than 70% of Nan in the person that are Poi and at total data set: 'director\_fees' (100%, 90%), 'restricted\_stock\_deferred' (100%, 88%), 'loan\_advances' (94%, 98%), 'deferral\_payments' (72%, 74%)

Same for 'deferred\_income' because that is a negative value that represent a discount in the payment and it is missing in 66% of the records

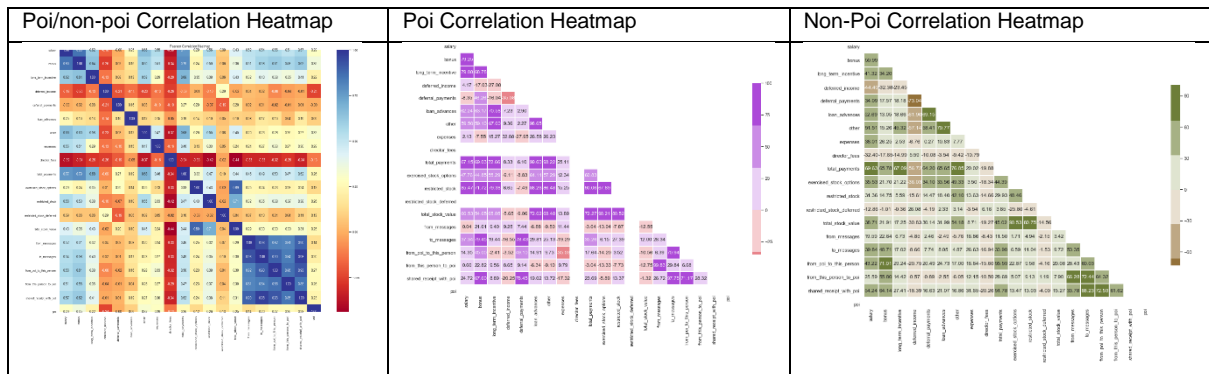
Regarding the emails, this project use the dataset corresponding to the counting of mails and not gone to the details of the text in the emails. Taking into account that the total amount of emails for each person has been cleaned in a more or less proper way, seams correct not to use that total amount but only the mails with relation with poi, that means send to or received from poi. On top, there are 40% of persons without mails

The level on Nan for poi and non-poi are high in the features we decide to remove from the data set. I remove 'deferred\_income' because that is a negative value that represent a discount in the payment and it is missing in 65% of the records (there off 70% for non-poi)

Regarding the emails, this project use the dataset corresponding to the counting of mails and not gone to the details of the text in the emails, we need to remember that the total amount of emails for each person has been cleaned in a more or less proper way, when the information that this indicator provide is limited and can be misleading. There is a big difference in level of Nan for poi and non-poi. All poi has emails address and only 22% have Nan for emails detailed features. When we check non-poi, the level of missing emails address is 26% and the emails detailed features Nan goes to 43% (nearly the half).

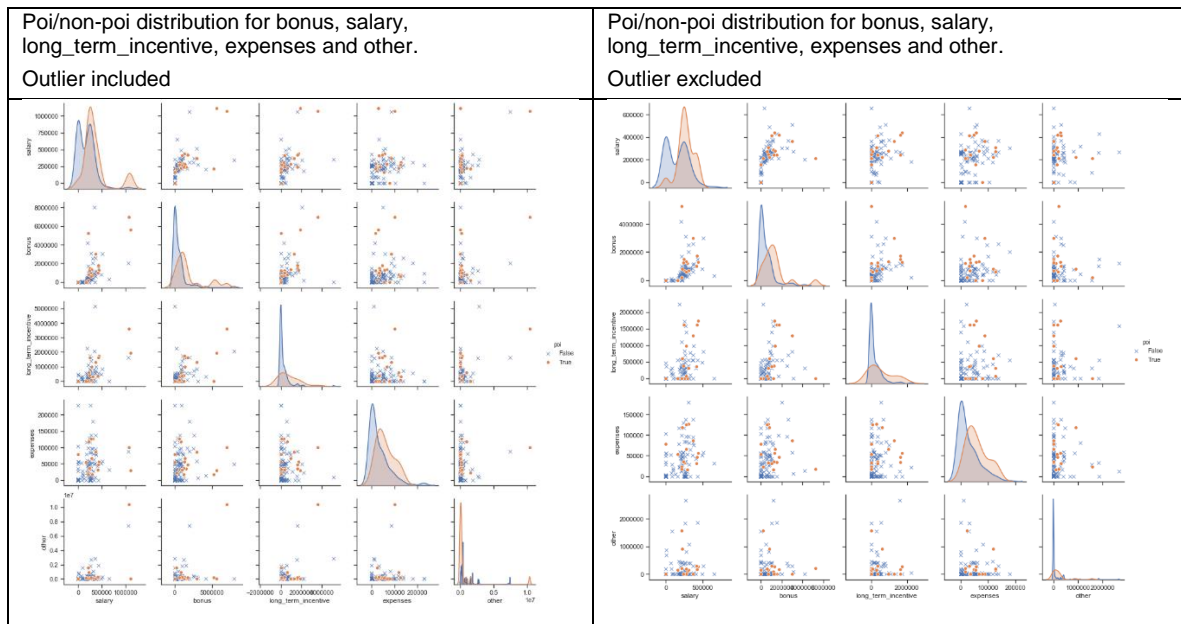
It is important to investigate the correlations between the features (see different approach and visualization in PML\_3\_Features\_analysis\_creation\_out.py)

Let's start with a global view for all features in a color map showing their correlation, for poi and non-poi



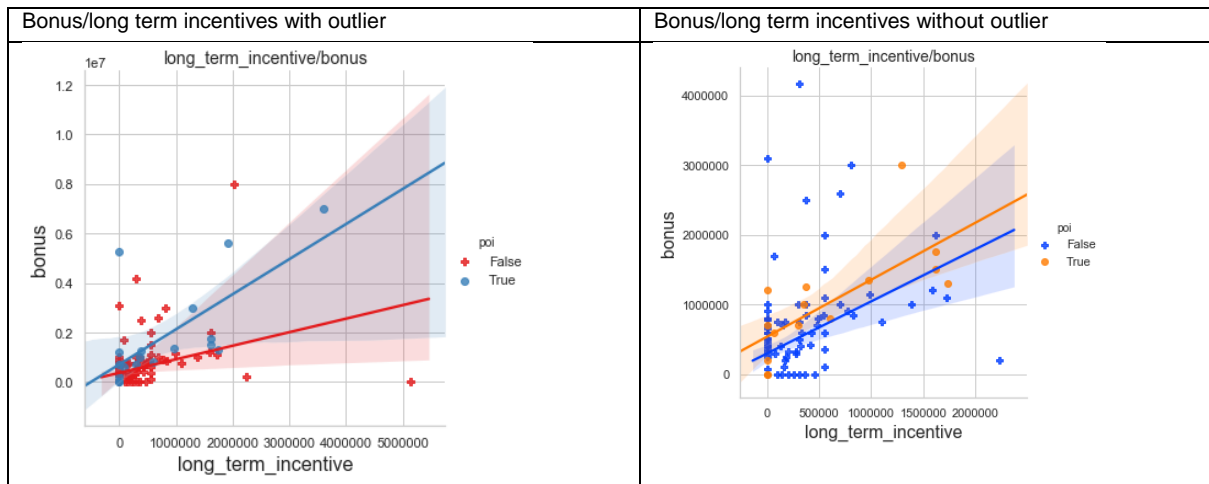
The features that have more correlation with all the other features are the ones that have cases with dark color in the in the heat-map: salary, bonus, long term incentives...

We can analyse the distribution of the different features for poi and non-poi, including and excluding outlier. Here below an example with features bonus, salary, long\_term\_incentive, expenses and other



We iterate this visual for the different features and we identify other outliers

Another visual done show the values split per poi and non-poi and the shadow of their tendency, that variate if we remove the outlier. Hereby one examples (see others in PLM\_3...)





We have now an idea of which features provides information valuable but the decisions of which features we pass to the algorithm could be enriched by several ways: creation of features, PCA use for creation of features, algorithm to help in selection (features importance, Select Kbest)

I should remove the features with big amount of Nan: loan\_advances (98%), director\_fees (90 %), restricted\_stock\_deferred (88%) and deferral\_payments (74%) as the added value information is minimum and the scaler will struggle with it. We will remove too 'deferred\_income'. Those features are visible in 3b\_Scaler\_features\_low\_entries

Next step is create new features. Since the moment we decide to remove features we need to find a way to keep that information if we consider it is relevant, a good way is to combine those features in a way that we could ensure that.

We can create features by combining others

In the case of payments and stock the new features we create will be focus in the extra payment the people get and its ratio over total money for the person:

"incentives" = bonus + long\_term\_incentive + exercised\_stock\_options

"total\_money" = total\_payments + total\_stock\_value;

We can use the new features to calculate ratios on the existing ones

"bonus\_ratio" = bonus / total\_money

"incentives\_ratio" = incentives / total\_money

Using same logic we could create several new features

"salary\_ratio" = (float( salary) / float( total\_money)

"expenses\_ratio" = (float( expenses) / float( total\_money)

"other\_ratio" = (float( other) / float( total\_money)

"long\_term\_incentive\_ratio" = (float( long\_term\_incentive) / float( total\_money)

"exercised\_stock\_options\_ratio" = (float( exercised\_stock\_options) / float( total\_money)

"total\_payments\_ratio" = (float( total\_payments) / float( total\_money)

"total\_stock\_value\_ratio" = (float(total\_stock\_value) / float( total\_money)

In the case of emails, the ratios show the relation of mails to poi and from poi regarding the amount of mails shared receipt with poi. If a poi send a mail to several people including other pois, this mail will be in from poi to this person and in shared receipt with poi.

"from\_poi\_to\_this\_person\_ratio" = (float( from\_poi\_to\_this\_person) / float( shared\_receipt\_with\_poi)

"from\_this\_person\_to\_poi\_ratio" = (float( from\_this\_person\_to\_poi) / float( shared\_receipt\_with\_poi)

We can create new features using PCA. The use of this technic is not so relevant in this case as we do not have too much features and on top the relations among them are not so strong but it is interesting to create that and see how it works.

payment\_f=['salary', 'bonus', 'long\_term\_incentive', 'other', 'expenses']

payment\_2=['salary', 'other', 'expenses']

payment\_tt=['salary', 'bonus', 'long\_term\_incentive', 'deferred\_income', 'deferral\_payments', 'loan\_advances', 'other', 'expenses', 'director\_fees']

incentive\_f=['bonus', 'long\_term\_incentive', 'exercised\_stock\_options']

emails\_exc=['from\_messages', 'to\_messages', 'from\_poi\_to\_this\_person', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi']

We have create 3 features regarding incentives

Sum of features @ "incentives" = bonus + long\_term\_incentive + exercised\_stock\_options

Ratio of features @ "incentives\_ratio" = incentives / total\_money

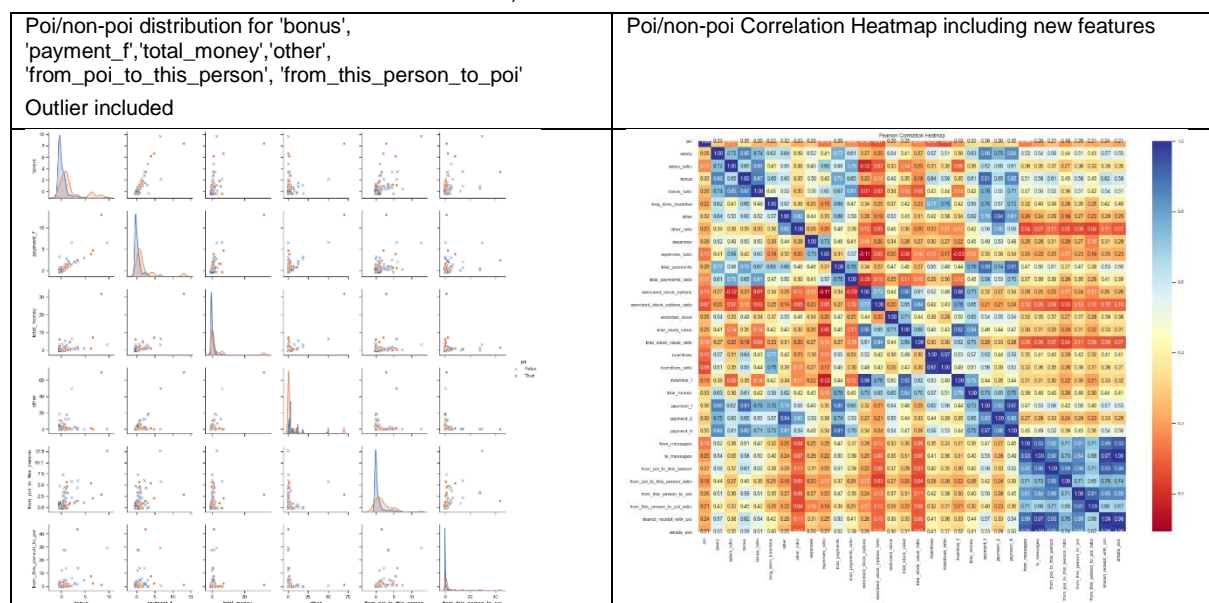
PCA @ incentive\_f=['bonus', 'long\_term\_incentive', 'exercised\_stock\_options']

The correlation of those features with poi and with the other features is different in each case, and their behavior with the algorithm are different too.

The spearman correlation of existing features is not so big, neither the one on the new features

Features spearman correlation with poi/non poi				Poi/non-poi Correlation Heatmap including new features																																																																																																																																																																																																																																																																																				
No features with more than 50% correlation, 9 features with correlation between 35% and 25% correlation				No features with more than 50% correlation, 14 features with correlation between 36% and 25% correlation.																																																																																																																																																																																																																																																																																				
<table><thead><tr><th></th><th>poi</th><th>total_payments</th><th>to_messages</th></tr></thead><tbody><tr><td>poi</td><td>100.0</td><td>28.0</td><td>20.0</td></tr><tr><td>bonus</td><td>35.0</td><td>79.0</td><td>58.0</td></tr><tr><td>other</td><td>32.0</td><td>69.0</td><td>24.0</td></tr><tr><td>salary</td><td>30.0</td><td>77.0</td><td>54.0</td></tr><tr><td>expenses</td><td>28.0</td><td>45.0</td><td>27.0</td></tr><tr><td>from_poi_to_this_person</td><td>27.0</td><td>50.0</td><td>90.0</td></tr><tr><td>from_this_person_to_poi</td><td>26.0</td><td>47.0</td><td>84.0</td></tr><tr><td>restricted_stock</td><td>26.0</td><td>46.0</td><td>35.0</td></tr><tr><td>total_payments</td><td>26.0</td><td>100.0</td><td>49.0</td></tr><tr><td>total_stock_value</td><td>25.0</td><td>44.0</td><td>31.0</td></tr><tr><td>shared_receipt_with_poi</td><td>24.0</td><td>52.0</td><td>98.0</td></tr><tr><td>long_term_incentive</td><td>22.0</td><td>68.0</td><td>40.0</td></tr><tr><td>to_messages</td><td>20.0</td><td>49.0</td><td>100.0</td></tr><tr><td>from_messages</td><td>14.0</td><td>46.0</td><td>93.0</td></tr><tr><td>restricted_stock_deferred</td><td>14.0</td><td>13.0</td><td>11.0</td></tr><tr><td>exercised_stock_options</td><td>12.0</td><td>32.0</td><td>25.0</td></tr><tr><td>loan_advances</td><td>9.0</td><td>19.0</td><td>17.0</td></tr><tr><td>deferral_payments</td><td>0.0</td><td>28.0</td><td>1.0</td></tr><tr><td>director_fees</td><td>-13.0</td><td>-35.0</td><td>-34.0</td></tr><tr><td>deferred_income</td><td>-24.0</td><td>-9.0</td><td>3.0</td></tr></tbody></table>					poi	total_payments	to_messages	poi	100.0	28.0	20.0	bonus	35.0	79.0	58.0	other	32.0	69.0	24.0	salary	30.0	77.0	54.0	expenses	28.0	45.0	27.0	from_poi_to_this_person	27.0	50.0	90.0	from_this_person_to_poi	26.0	47.0	84.0	restricted_stock	26.0	46.0	35.0	total_payments	26.0	100.0	49.0	total_stock_value	25.0	44.0	31.0	shared_receipt_with_poi	24.0	52.0	98.0	long_term_incentive	22.0	68.0	40.0	to_messages	20.0	49.0	100.0	from_messages	14.0	46.0	93.0	restricted_stock_deferred	14.0	13.0	11.0	exercised_stock_options	12.0	32.0	25.0	loan_advances	9.0	19.0	17.0	deferral_payments	0.0	28.0	1.0	director_fees	-13.0	-35.0	-34.0	deferred_income	-24.0	-9.0	3.0	<table><thead><tr><th></th><th>poi</th><th>incentives</th><th>incentives_ratio</th><th>incentive_f</th></tr></thead><tbody><tr><td>poi</td><td>100.0</td><td>10.0</td><td>6.0</td><td>19.0</td></tr><tr><td>payment_f</td><td>36.0</td><td>61.0</td><td>55.0</td><td>43.0</td></tr><tr><td>bonus</td><td>35.0</td><td>64.0</td><td>56.0</td><td>35.0</td></tr><tr><td>payment_tt</td><td>34.0</td><td>58.0</td><td>52.0</td><td>42.0</td></tr><tr><td>total_money</td><td>33.0</td><td>57.0</td><td>51.0</td><td>79.0</td></tr><tr><td>other</td><td>32.0</td><td>43.0</td><td>39.0</td><td>34.0</td></tr><tr><td>salary</td><td>30.0</td><td>57.0</td><td>51.0</td><td>39.0</td></tr><tr><td>payment_2</td><td>29.0</td><td>44.0</td><td>38.0</td><td>33.0</td></tr><tr><td>expenses</td><td>28.0</td><td>30.0</td><td>27.0</td><td>21.0</td></tr><tr><td>from_poi_to_this_person</td><td>27.0</td><td>40.0</td><td>35.0</td><td>30.0</td></tr><tr><td>from_this_person_to_poi</td><td>26.0</td><td>42.0</td><td>38.0</td><td>30.0</td></tr><tr><td>bonus_ratio</td><td>26.0</td><td>43.0</td><td>44.0</td><td>14.0</td></tr><tr><td>restricted_stock</td><td>26.0</td><td>38.0</td><td>29.0</td><td>50.0</td></tr><tr><td>total_payments</td><td>26.0</td><td>54.0</td><td>49.0</td><td>43.0</td></tr><tr><td>total_stock_value</td><td>25.0</td><td>49.0</td><td>43.0</td><td>92.0</td></tr><tr><td>shared_receipt_with_poi</td><td>24.0</td><td>41.0</td><td>35.0</td><td>33.0</td></tr><tr><td>other_ratio</td><td>23.0</td><td>23.0</td><td>17.0</td><td>17.0</td></tr><tr><td>long_term_incentive</td><td>22.0</td><td>77.0</td><td>75.0</td><td>42.0</td></tr><tr><td>from_this_person_to_poi_ratio</td><td>21.0</td><td>32.0</td><td>31.0</td><td>21.0</td></tr><tr><td>emails_exc</td><td>21.0</td><td>41.0</td><td>37.0</td><td>32.0</td></tr><tr><td>to_messages</td><td>20.0</td><td>41.0</td><td>36.0</td><td>32.0</td></tr><tr><td>incentive_f</td><td>19.0</td><td>53.0</td><td>48.0</td><td>100.0</td></tr><tr><td>from_poi_to_this_person_ratio</td><td>18.0</td><td>28.0</td><td>26.0</td><td>23.0</td></tr><tr><td>total_payments_ratio</td><td>17.0</td><td>29.0</td><td>30.0</td><td>12.0</td></tr><tr><td>salary_ratio</td><td>17.0</td><td>31.0</td><td>35.0</td><td>9.0</td></tr><tr><td>total_stock_value_ratio</td><td>16.0</td><td>30.0</td><td>30.0</td><td>62.0</td></tr><tr><td>from_messages</td><td>14.0</td><td>35.0</td><td>33.0</td><td>31.0</td></tr><tr><td>restricted_stock_deferred</td><td>14.0</td><td>-1.0</td><td>-3.0</td><td>-3.0</td></tr><tr><td>expenses_ratio</td><td>13.0</td><td>15.0</td><td>17.0</td><td>-3.0</td></tr><tr><td>exercised_stock_options</td><td>12.0</td><td>52.0</td><td>48.0</td><td>98.0</td></tr><tr><td>incentives</td><td>10.0</td><td>100.0</td><td>97.0</td><td>53.0</td></tr><tr><td>loan_advances</td><td>6.0</td><td>17.0</td><td>5.0</td><td>12.0</td></tr><tr><td>exercised_stock_options_ratio</td><td>7.0</td><td>42.0</td><td>43.0</td><td>76.0</td></tr><tr><td>incentives_ratio</td><td>6.0</td><td>97.0</td><td>100.0</td><td>48.0</td></tr><tr><td>deferral_payments</td><td>0.0</td><td>12.0</td><td>6.0</td><td>28.0</td></tr><tr><td>director_fees</td><td>-13.0</td><td>-22.0</td><td>-22.0</td><td>-42.0</td></tr><tr><td>deferred_income</td><td>-24.0</td><td>-13.0</td><td>-10.0</td><td>1.0</td></tr></tbody></table>				poi	incentives	incentives_ratio	incentive_f	poi	100.0	10.0	6.0	19.0	payment_f	36.0	61.0	55.0	43.0	bonus	35.0	64.0	56.0	35.0	payment_tt	34.0	58.0	52.0	42.0	total_money	33.0	57.0	51.0	79.0	other	32.0	43.0	39.0	34.0	salary	30.0	57.0	51.0	39.0	payment_2	29.0	44.0	38.0	33.0	expenses	28.0	30.0	27.0	21.0	from_poi_to_this_person	27.0	40.0	35.0	30.0	from_this_person_to_poi	26.0	42.0	38.0	30.0	bonus_ratio	26.0	43.0	44.0	14.0	restricted_stock	26.0	38.0	29.0	50.0	total_payments	26.0	54.0	49.0	43.0	total_stock_value	25.0	49.0	43.0	92.0	shared_receipt_with_poi	24.0	41.0	35.0	33.0	other_ratio	23.0	23.0	17.0	17.0	long_term_incentive	22.0	77.0	75.0	42.0	from_this_person_to_poi_ratio	21.0	32.0	31.0	21.0	emails_exc	21.0	41.0	37.0	32.0	to_messages	20.0	41.0	36.0	32.0	incentive_f	19.0	53.0	48.0	100.0	from_poi_to_this_person_ratio	18.0	28.0	26.0	23.0	total_payments_ratio	17.0	29.0	30.0	12.0	salary_ratio	17.0	31.0	35.0	9.0	total_stock_value_ratio	16.0	30.0	30.0	62.0	from_messages	14.0	35.0	33.0	31.0	restricted_stock_deferred	14.0	-1.0	-3.0	-3.0	expenses_ratio	13.0	15.0	17.0	-3.0	exercised_stock_options	12.0	52.0	48.0	98.0	incentives	10.0	100.0	97.0	53.0	loan_advances	6.0	17.0	5.0	12.0	exercised_stock_options_ratio	7.0	42.0	43.0	76.0	incentives_ratio	6.0	97.0	100.0	48.0	deferral_payments	0.0	12.0	6.0	28.0	director_fees	-13.0	-22.0	-22.0	-42.0	deferred_income	-24.0	-13.0	-10.0	1.0
	poi	total_payments	to_messages																																																																																																																																																																																																																																																																																					
poi	100.0	28.0	20.0																																																																																																																																																																																																																																																																																					
bonus	35.0	79.0	58.0																																																																																																																																																																																																																																																																																					
other	32.0	69.0	24.0																																																																																																																																																																																																																																																																																					
salary	30.0	77.0	54.0																																																																																																																																																																																																																																																																																					
expenses	28.0	45.0	27.0																																																																																																																																																																																																																																																																																					
from_poi_to_this_person	27.0	50.0	90.0																																																																																																																																																																																																																																																																																					
from_this_person_to_poi	26.0	47.0	84.0																																																																																																																																																																																																																																																																																					
restricted_stock	26.0	46.0	35.0																																																																																																																																																																																																																																																																																					
total_payments	26.0	100.0	49.0																																																																																																																																																																																																																																																																																					
total_stock_value	25.0	44.0	31.0																																																																																																																																																																																																																																																																																					
shared_receipt_with_poi	24.0	52.0	98.0																																																																																																																																																																																																																																																																																					
long_term_incentive	22.0	68.0	40.0																																																																																																																																																																																																																																																																																					
to_messages	20.0	49.0	100.0																																																																																																																																																																																																																																																																																					
from_messages	14.0	46.0	93.0																																																																																																																																																																																																																																																																																					
restricted_stock_deferred	14.0	13.0	11.0																																																																																																																																																																																																																																																																																					
exercised_stock_options	12.0	32.0	25.0																																																																																																																																																																																																																																																																																					
loan_advances	9.0	19.0	17.0																																																																																																																																																																																																																																																																																					
deferral_payments	0.0	28.0	1.0																																																																																																																																																																																																																																																																																					
director_fees	-13.0	-35.0	-34.0																																																																																																																																																																																																																																																																																					
deferred_income	-24.0	-9.0	3.0																																																																																																																																																																																																																																																																																					
	poi	incentives	incentives_ratio	incentive_f																																																																																																																																																																																																																																																																																				
poi	100.0	10.0	6.0	19.0																																																																																																																																																																																																																																																																																				
payment_f	36.0	61.0	55.0	43.0																																																																																																																																																																																																																																																																																				
bonus	35.0	64.0	56.0	35.0																																																																																																																																																																																																																																																																																				
payment_tt	34.0	58.0	52.0	42.0																																																																																																																																																																																																																																																																																				
total_money	33.0	57.0	51.0	79.0																																																																																																																																																																																																																																																																																				
other	32.0	43.0	39.0	34.0																																																																																																																																																																																																																																																																																				
salary	30.0	57.0	51.0	39.0																																																																																																																																																																																																																																																																																				
payment_2	29.0	44.0	38.0	33.0																																																																																																																																																																																																																																																																																				
expenses	28.0	30.0	27.0	21.0																																																																																																																																																																																																																																																																																				
from_poi_to_this_person	27.0	40.0	35.0	30.0																																																																																																																																																																																																																																																																																				
from_this_person_to_poi	26.0	42.0	38.0	30.0																																																																																																																																																																																																																																																																																				
bonus_ratio	26.0	43.0	44.0	14.0																																																																																																																																																																																																																																																																																				
restricted_stock	26.0	38.0	29.0	50.0																																																																																																																																																																																																																																																																																				
total_payments	26.0	54.0	49.0	43.0																																																																																																																																																																																																																																																																																				
total_stock_value	25.0	49.0	43.0	92.0																																																																																																																																																																																																																																																																																				
shared_receipt_with_poi	24.0	41.0	35.0	33.0																																																																																																																																																																																																																																																																																				
other_ratio	23.0	23.0	17.0	17.0																																																																																																																																																																																																																																																																																				
long_term_incentive	22.0	77.0	75.0	42.0																																																																																																																																																																																																																																																																																				
from_this_person_to_poi_ratio	21.0	32.0	31.0	21.0																																																																																																																																																																																																																																																																																				
emails_exc	21.0	41.0	37.0	32.0																																																																																																																																																																																																																																																																																				
to_messages	20.0	41.0	36.0	32.0																																																																																																																																																																																																																																																																																				
incentive_f	19.0	53.0	48.0	100.0																																																																																																																																																																																																																																																																																				
from_poi_to_this_person_ratio	18.0	28.0	26.0	23.0																																																																																																																																																																																																																																																																																				
total_payments_ratio	17.0	29.0	30.0	12.0																																																																																																																																																																																																																																																																																				
salary_ratio	17.0	31.0	35.0	9.0																																																																																																																																																																																																																																																																																				
total_stock_value_ratio	16.0	30.0	30.0	62.0																																																																																																																																																																																																																																																																																				
from_messages	14.0	35.0	33.0	31.0																																																																																																																																																																																																																																																																																				
restricted_stock_deferred	14.0	-1.0	-3.0	-3.0																																																																																																																																																																																																																																																																																				
expenses_ratio	13.0	15.0	17.0	-3.0																																																																																																																																																																																																																																																																																				
exercised_stock_options	12.0	52.0	48.0	98.0																																																																																																																																																																																																																																																																																				
incentives	10.0	100.0	97.0	53.0																																																																																																																																																																																																																																																																																				
loan_advances	6.0	17.0	5.0	12.0																																																																																																																																																																																																																																																																																				
exercised_stock_options_ratio	7.0	42.0	43.0	76.0																																																																																																																																																																																																																																																																																				
incentives_ratio	6.0	97.0	100.0	48.0																																																																																																																																																																																																																																																																																				
deferral_payments	0.0	12.0	6.0	28.0																																																																																																																																																																																																																																																																																				
director_fees	-13.0	-22.0	-22.0	-42.0																																																																																																																																																																																																																																																																																				
deferred_income	-24.0	-13.0	-10.0	1.0																																																																																																																																																																																																																																																																																				

Out of the 19 features we have created, 5 have a correlation over 25%.



All features have low correlation level with poi, but some features has high correlation among them. That means we can reduce the features taking into account the ones that are correlated among them and low with the poi.

We can see those features on PML\_3\_features\_analysis\_creation.ipynb.

We will use some of them depending on their importance for the algorithm.

The algorithm to determine feature importance will work only if the features we pass has a sense, it means if our data quality is bad, it could lead to a bias that provide result without any sense.

The objective of this algorithm is to classify the entries in two groups (poi, non\_poi). We need to use classification algorithm instead of regression. SVM or Kmeans are good examples, both of them need feature scaling. We need to scale the features are we are using different units (\$, number of emails, ratio) to ensure the classification algorithm understand the information properly.

There are several scalar, I tested MinMaxScaler and RobustScaler.

We pass the features over those scalar to see what it does to them. That is visible in the table below

Features describe before scaling									Features after Robust scaler									Features after MinMax scaler								
	count	mean	std	min	25%	50%	75%	max		count	mean	std	min	25%	50%	75%	max		count	mean	std	min	25%	50%	75%	max
total_money	144.0	5004016.0	1379577.0	0.0	0.0	2008894.0	4706047.0	15266871.0	payment_1	144.0	5.0	58.0	-0.0	-0.0	0.0	1.0	791.0	bonus	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
total_payments	144.0	2149476.0	8779364.0	0.0	87472.0	913825.0	1885158.0	10359793.0	from_messages	144.0	7.0	27.0	-0.0	-0.0	0.0	1.0	271.0	bonus_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
payment_1	144.0	-0.0	6876981.0	-684482.0	-684128.0	-631654.0	-566446.0	61878820.0	other	144.0	2.0	8.0	-0.0	-0.0	0.0	1.0	79.0	deferred_income	144.0	-0.0	0.0	-1.0	-0.0	0.0	0.0	0.0
total_stock_value	144.0	3017427.0	6271528.0	0.0	256376.0	988534.0	2372703.0	49110078.0	payment_2	144.0	2.0	7.0	-0.0	-0.0	0.0	1.0	63.0	emails_exc	144.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	0.0
incentives	144.0	1860333.0	5253559.0	0.0	0.0	0.0	1587674.0	44948384.0	from_this_person_to_poi	144.0	2.0	6.0	0.0	0.0	0.0	1.0	44.0	exercised_stock_options	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
incentive_f	144.0	-0.0	4956138.0	-2239062.0	-2158882.0	-1573413.0	-493664.0	32798537.0	total_payments	144.0	1.0	5.0	-1.0	-0.0	0.0	1.0	57.0	exercised_stock_options_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
exercised_stock_options	144.0	2155028.0	4923320.0	0.0	0.0	608294.0	1683580.0	34438384.0	from_this_person_to_poi_ratio	144.0	2.0	5.0	0.0	0.0	0.0	1.0	33.0	expenses	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
restricted_stock	144.0	905018.0	2000387.0	0.0	44093.0	361678.0	857103.0	14761694.0	total_stock_value	144.0	1.0	3.0	-0.0	-0.0	0.0	1.0	23.0	expenses_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
payment_f	144.0	0.0	1475715.0	-794562.0	-794552.0	-445268.0	143438.0	11762310.0	exercised_stock_options	144.0	1.0	3.0	-0.0	-0.0	0.0	1.0	20.0	from_messages	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
bonus	144.0	675997.0	123315.0	0.0	0.0	300000.0	800000.0	8000000.0	total_money	144.0	1.0	3.0	-0.0	-0.0	-0.0	1.0	32.0	from_poi_to_this_person	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
payment_2	144.0	-0.0	1136736.0	-310521.0	-310515.0	-284723.0	-144575.0	10104736.0	incentive_f	144.0	1.0	3.0	-0.0	-0.0	0.0	1.0	21.0	from_poi_to_this_person_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
other	144.0	293788.0	1131517.0	0.0	0.0	882.0	148577.0	10399726.0	incentives	144.0	1.0	3.0	0.0	0.0	0.0	1.0	20.0	from_this_person_to_poi	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
long_term_incentive	144.0	336958.0	687183.0	0.0	0.0	0.0	374588.0	5145434.0	other_ratio	144.0	2.0	3.0	0.0	0.0	0.0	1.0	16.0	from_this_person_to_poi_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
salary	144.0	185446.0	197042.0	0.0	0.0	216996.0	269668.0	1111258.0	bonus	144.0	0.0	2.0	-0.0	-0.0	0.0	1.0	10.0	incentive_f	144.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	1.0
emails	144.0	36356.0	49990.0	0.0	0.0	21937.0	54234.0	228763.0	restricted_stock	144.0	1.0	2.0	-0.0	-0.0	-0.0	1.0	18.0	incentives	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
emails_exc	144.0	0.0	2569.0	-1453.0	-1453.0	-1062.0	440.0	13665.0	payment_f	144.0	0.0	2.0	-0.0	-0.0	0.0	1.0	13.0	incentive_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
to_messages	144.0	1239.0	2238.0	0.0	0.0	340.0	1623.0	15169.0	long_term_incentive	144.0	0.0	2.0	0.0	0.0	0.0	1.0	14.0	long_term_incentive	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
from_messages	144.0	364.0	1451.0	0.0	0.0	18.0	53.0	14368.0	other	144.0	2.0	-0.0	-0.0	-0.0	0.0	1.0	13.0	other	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
shared_receipt_with_poi	144.0	793.0	1077.0	0.0	0.0	114.0	934.0	5921.0	bonus_ratio	144.0	0.0	1.0	-0.0	-0.0	0.0	1.0	3.0	other_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
from_this_person_to_poi	144.0	25.0	80.0	0.0	0.0	0.0	14.0	609.0	incentives_ratio	144.0	0.0	1.0	0.0	0.0	0.0	1.0	2.0	payment_2	144.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	1.0
from_poi_to_this_person	144.0	39.0	74.0	0.0	0.0	0.0	4.0	41.0	payment_f	144.0	-0.0	0.0	-0.0	-0.0	-0.0	0.0	6.0	payment_1	144.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	1.0
incentives_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	payment_2	144.0	-0.0	0.0	-0.0	-0.0	-0.0	-0.0	11.0	restricted_stock	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
total_payments_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	salary	144.0	-0.0	1.0	-1.0	-1.0	0.0	0.0	3.0	salary	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
salary_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	shared_receipt_with_poi	144.0	1.0	1.0	-0.0	-0.0	0.0	1.0	6.0	shared_receipt_with_poi	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
from_poi_to_this_person_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	to_messages	144.0	1.0	1.0	-0.0	-0.0	0.0	1.0	9.0	to_messages	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
other_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	expenses	144.0	0.0	1.0	-0.0	-0.0	0.0	1.0	4.0	total_money	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
bonus_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	exercised_stock_options_ratio	144.0	0.0	1.0	-0.0	-0.0	0.0	1.0	2.0	total_payments_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
total_stock_value_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0	emails_exc	144.0	1.0	1.0	-0.0	-0.0	0.0	1.0	8.0	total_payments	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
expenses_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	total_stock_value_ratio	144.0	0.0	1.0	-1.0	-1.0	0.0	0.0	1.0	total_stock_value_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0
exercised_stock_options_ratio	144.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	total_payments_ratio	144.0	0.0	0.0	-1.0	-1.0	0.0	0.0	1.0	total_stock_value	144.0	0.0	0.0	0.0	0.0	0.0	0.0	1.0

We see that the information we pass to the algorithm will be very different depending on using or not scaler and which scaler we use. I will use Robust scaler as the data have outlier and we need to keep that information. Robust scaler use the median and the interquartile range to scale each feature. Further details are available at 3a [Scaler\\_features.ipynb](#) and 3b [Scaler\\_features](#)

First we need to scale the features as the features are different units (money, amount of mails and ratios) I will use RobustScaler as the data have outlier and we need to keep that information.

The correlation among the features does not change after scaling. The correlations between the features and the poi/non poi is not significant for any features that means the algorithm is going to struggle to find the right solution.

We need to use the right number of features and the right features to get the algorithm to predict in the best way.

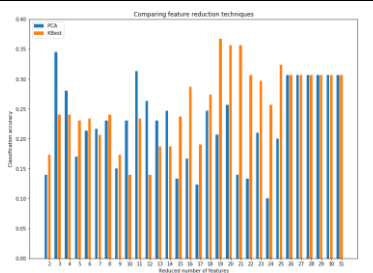
The main issue with the features we have is that they have a very low linear correlation with the independent variable. That will difficult to reduce the number of features to use

The first features reduction is the one done before when we remove all the features with more than 70% of Nan in the person that are Poi: 'director\_fees', 'restricted\_stock\_deferred', 'loan\_advances', 'deferral\_payments' and 'deferred\_income'

The features selection could be done independent of the algorithm using KBest, PCA, for example, or integrate in a Pipe and run it at same time than the algorithm

The point is that the importance of the features depend on the algorithm as they operate differently and then the information extract form the features is different.

And the selector will operate differently depending on the features scaled method is used.

Classifier: AdaBoost (n_estimators=50, random_state=42)], Reduction features algorithm and number of features		
Features selection before scaling	Features selection after Robust scaler	Features selection after MinMax scaler
 <p>PCA: n_components=16</p>	 <p>SelectKBest(k=19: ['salary', 'payment_tt', 'payment_f', 'restricted_stock', 'shared_receipt_with_poi', 'incentives', 'incentive_f', 'bonus', 'total_payments_ratio', 'bonus_ratio', 'long_term_incentive', 'total_money', 'total_stock_value', 'total_payments', 'from_this_person_to_poi', 'from_poi_to_this_person', 'exercised_stock_options', 'expenses', 'exercised_stock_options_ratio'])</p>	 <p>SelectKBest(k=19: ['salary', 'payment_f', 'restricted_stock', 'shared_receipt_with_poi', 'incentives', 'incentive_f', 'bonus', 'total_stock_value_ratio', 'total_payments_ratio', 'bonus_ratio', 'long_term_incentive', 'total_money', 'total_stock_value', 'total_payments', 'from_this_person_to_poi', 'from_poi_to_this_person', 'exercised_stock_options', 'expenses', 'exercised_stock_options_ratio'])</p>

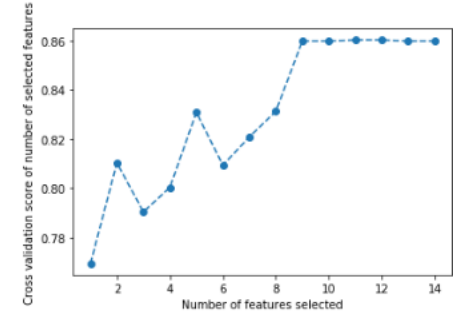
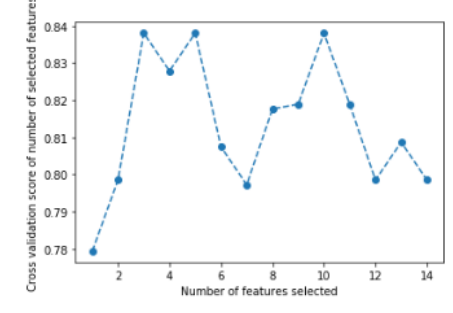
Detailed analysis of different features including and excluding outlier are available on [PML\\_3\\_features\\_analysis\\_creation.ipynb](#) and further details on scaling are available at [3a\\_Scaler\\_features.ipynb](#). and [3b\\_Scaler\\_features\\_low\\_amount\\_entries.ipynb](#)

The idea of create new features is to be able to provide same information to the algorithm with less features but in our case, the number of features is limited, then not really need to use.

The selection of the features has been done once the algorithm to use has been decided.

We pass different features to the algorithm at same time that we fine-tune it to get the best combination.

We use RFECV to estimate the best number of features and we compare with KBest result

RandomForestClassifier(random_state=42)	Pipeline with KBest
<pre> forest = RandomForestClassifier(random_state=42) rfecv = RFECV(estimator=forest, cv=StratifiedKFold(5), scoring='accuracy') rfecv = rfecv.fit(X_train, y_train) </pre> 	<pre> pipe=Pipeline([('select',SelectKBest(k=9)),('rfc', RandomForestClassifier (random_state=42))]) pipe.fit(features_test,labels_test) pipe.score(features_test,labels_test) </pre> <p>0.9545454545454546</p>
AdaBoost (random_state=42)	Pipeline with KBest
<pre> clf_ada=AdaBoostClassifier()# (random_state=42) rfecv = RFECV(estimator=clf_ada, cv=14, scoring='accuracy') rfecv = rfecv.fit(X_train, y_train) </pre> 	<pre> pipe=Pipeline([('select',SelectKBest(k=10)),('Ada', AdaBoostClassifier())]) #f_classif,k=4 SelectFdr pipe.fit(X,y) pipe.score(features_test,labels_test) </pre> <p>0.8863636363636364</p>

RandomForestClassifier(random_state=42) with different features		Features importance																																																																											
<pre>In [9]: 1 clf_rf=RandomForestClassifier(random_state=42) 2 3 pass_clf(algo_list0, features_list1, output_list0, data_df, clf_rf)</pre> <div>CONFUSION MATRIX</div> <pre>[[40  0]  [ 2 21]]</pre> <div>CLASSIFICATION REPORT</div> <table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>False</td><td>0.95</td><td>1.00</td><td>0.98</td><td>40</td></tr><tr><td>True</td><td>1.00</td><td>0.50</td><td>0.67</td><td>4</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.95</td><td>44</td></tr><tr><td>macro avg</td><td>0.98</td><td>0.75</td><td>0.82</td><td>44</td></tr><tr><td>weighted avg</td><td>0.96</td><td>0.95</td><td>0.95</td><td>44</td></tr></tbody></table>			precision	recall	f1-score	support	False	0.95	1.00	0.98	40	True	1.00	0.50	0.67	4	accuracy			0.95	44	macro avg	0.98	0.75	0.82	44	weighted avg	0.96	0.95	0.95	44	<table><thead><tr><th></th><th>feature</th><th>importance</th></tr></thead><tbody><tr><td>10</td><td>total_stock_value</td><td>0.163088</td></tr><tr><td>7</td><td>expenses</td><td>0.141091</td></tr><tr><td>5</td><td>bonus</td><td>0.104008</td></tr><tr><td>9</td><td>long_term_incentive</td><td>0.103548</td></tr><tr><td>4</td><td>other</td><td>0.091039</td></tr><tr><td>11</td><td>exercised_stock_options</td><td>0.079954</td></tr><tr><td>12</td><td>from_this_person_to_poi</td><td>0.065708</td></tr><tr><td>2</td><td>total_payments</td><td>0.061876</td></tr><tr><td>8</td><td>from_poi_to_this_person</td><td>0.046845</td></tr><tr><td>1</td><td>from_messages</td><td>0.045101</td></tr><tr><td>0</td><td>to_messages</td><td>0.035342</td></tr><tr><td>3</td><td>restricted_stock</td><td>0.030596</td></tr><tr><td>6</td><td>salary</td><td>0.019653</td></tr><tr><td>13</td><td>shared_receipt_with_poi</td><td>0.012152</td></tr></tbody></table>		feature	importance	10	total_stock_value	0.163088	7	expenses	0.141091	5	bonus	0.104008	9	long_term_incentive	0.103548	4	other	0.091039	11	exercised_stock_options	0.079954	12	from_this_person_to_poi	0.065708	2	total_payments	0.061876	8	from_poi_to_this_person	0.046845	1	from_messages	0.045101	0	to_messages	0.035342	3	restricted_stock	0.030596	6	salary	0.019653	13	shared_receipt_with_poi	0.012152
	precision	recall	f1-score	support																																																																									
False	0.95	1.00	0.98	40																																																																									
True	1.00	0.50	0.67	4																																																																									
accuracy			0.95	44																																																																									
macro avg	0.98	0.75	0.82	44																																																																									
weighted avg	0.96	0.95	0.95	44																																																																									
	feature	importance																																																																											
10	total_stock_value	0.163088																																																																											
7	expenses	0.141091																																																																											
5	bonus	0.104008																																																																											
9	long_term_incentive	0.103548																																																																											
4	other	0.091039																																																																											
11	exercised_stock_options	0.079954																																																																											
12	from_this_person_to_poi	0.065708																																																																											
2	total_payments	0.061876																																																																											
8	from_poi_to_this_person	0.046845																																																																											
1	from_messages	0.045101																																																																											
0	to_messages	0.035342																																																																											
3	restricted_stock	0.030596																																																																											
6	salary	0.019653																																																																											
13	shared_receipt_with_poi	0.012152																																																																											
<pre>1 algo_list_rf=[ 'total_stock_value','expenses', 'bonus','long_term_incentive', 2               'other','exercised_stock_options','from_this_person_to_poi', 3               'total_payments','from_poi_to_this_person' 4               ] 5 6 clf_rf=RandomForestClassifier (random_state=42) 7 8 features_list_rf=algo_list_rf.copy() 9 features_list_rf.insert(0,'poi') 10 11 output_list_rf=features_list_rf.copy() 12 output_list_rf.insert(1,'poi_pred') 13 14 pass_clf(algo_list_rf, features_list_rf, output_list_rf, data_df, clf_rf)</pre> <div>CONFUSION MATRIX</div> <pre>[[37  3]  [ 2 21]]</pre> <div>CLASSIFICATION REPORT</div> <table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>False</td><td>0.95</td><td>0.93</td><td>0.94</td><td>40</td></tr><tr><td>True</td><td>0.40</td><td>0.50</td><td>0.44</td><td>4</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.89</td><td>44</td></tr><tr><td>macro avg</td><td>0.67</td><td>0.71</td><td>0.69</td><td>44</td></tr><tr><td>weighted avg</td><td>0.90</td><td>0.89</td><td>0.89</td><td>44</td></tr></tbody></table>			precision	recall	f1-score	support	False	0.95	0.93	0.94	40	True	0.40	0.50	0.44	4	accuracy			0.89	44	macro avg	0.67	0.71	0.69	44	weighted avg	0.90	0.89	0.89	44	<div>PRECISION</div> <p>0.4</p> <div>RECALL</div> <p>0.5</p> <table><thead><tr><th></th><th>feature</th><th>importance</th></tr></thead><tbody><tr><td>1</td><td>expenses</td><td>0.206501</td></tr><tr><td>5</td><td>exercised_stock_options</td><td>0.168430</td></tr><tr><td>0</td><td>total_stock_value</td><td>0.140569</td></tr><tr><td>6</td><td>from_this_person_to_poi</td><td>0.112185</td></tr><tr><td>2</td><td>bonus</td><td>0.110918</td></tr><tr><td>7</td><td>total_payments</td><td>0.092223</td></tr><tr><td>4</td><td>other</td><td>0.060570</td></tr><tr><td>8</td><td>from_poi_to_this_person</td><td>0.056828</td></tr><tr><td>3</td><td>long_term_incentive</td><td>0.051775</td></tr></tbody></table>		feature	importance	1	expenses	0.206501	5	exercised_stock_options	0.168430	0	total_stock_value	0.140569	6	from_this_person_to_poi	0.112185	2	bonus	0.110918	7	total_payments	0.092223	4	other	0.060570	8	from_poi_to_this_person	0.056828	3	long_term_incentive	0.051775															
	precision	recall	f1-score	support																																																																									
False	0.95	0.93	0.94	40																																																																									
True	0.40	0.50	0.44	4																																																																									
accuracy			0.89	44																																																																									
macro avg	0.67	0.71	0.69	44																																																																									
weighted avg	0.90	0.89	0.89	44																																																																									
	feature	importance																																																																											
1	expenses	0.206501																																																																											
5	exercised_stock_options	0.168430																																																																											
0	total_stock_value	0.140569																																																																											
6	from_this_person_to_poi	0.112185																																																																											
2	bonus	0.110918																																																																											
7	total_payments	0.092223																																																																											
4	other	0.060570																																																																											
8	from_poi_to_this_person	0.056828																																																																											
3	long_term_incentive	0.051775																																																																											
<pre>1 algo_list_rf=[ 'exercised_stock_options','other','bonus','from_this_person_to_poi','total_payments' 2               ] 3 clf_rf=RandomForestClassifier (random_state=42) 4 5 features_list_rf=algo_list_rf.copy() 6 features_list_rf.insert(0,'poi') 7 8 output_list_rf=features_list_rf.copy() 9 output_list_rf.insert(1,'poi_pred') 10 11 pass_clf(algo_list_rf, features_list_rf, output_list_rf, data_df, clf_rf)</pre> <div>CONFUSION MATRIX</div> <pre>[[39  1]  [ 1 31]]</pre> <div>CLASSIFICATION REPORT</div> <table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>False</td><td>0.97</td><td>0.97</td><td>0.97</td><td>40</td></tr><tr><td>True</td><td>0.75</td><td>0.75</td><td>0.75</td><td>4</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.95</td><td>44</td></tr><tr><td>macro avg</td><td>0.86</td><td>0.86</td><td>0.86</td><td>44</td></tr><tr><td>weighted avg</td><td>0.95</td><td>0.95</td><td>0.95</td><td>44</td></tr></tbody></table>			precision	recall	f1-score	support	False	0.97	0.97	0.97	40	True	0.75	0.75	0.75	4	accuracy			0.95	44	macro avg	0.86	0.86	0.86	44	weighted avg	0.95	0.95	0.95	44	<div>PRECISION</div> <p>0.75</p> <div>RECALL</div> <p>0.75</p> <table><thead><tr><th></th><th>feature</th><th>importance</th></tr></thead><tbody><tr><td>0</td><td>exercised_stock_options</td><td>0.401301</td></tr><tr><td>4</td><td>total_payments</td><td>0.200014</td></tr><tr><td>2</td><td>bonus</td><td>0.193352</td></tr><tr><td>1</td><td>other</td><td>0.106484</td></tr><tr><td>3</td><td>from_this_person_to_poi</td><td>0.098848</td></tr></tbody></table>		feature	importance	0	exercised_stock_options	0.401301	4	total_payments	0.200014	2	bonus	0.193352	1	other	0.106484	3	from_this_person_to_poi	0.098848																											
	precision	recall	f1-score	support																																																																									
False	0.97	0.97	0.97	40																																																																									
True	0.75	0.75	0.75	4																																																																									
accuracy			0.95	44																																																																									
macro avg	0.86	0.86	0.86	44																																																																									
weighted avg	0.95	0.95	0.95	44																																																																									
	feature	importance																																																																											
0	exercised_stock_options	0.401301																																																																											
4	total_payments	0.200014																																																																											
2	bonus	0.193352																																																																											
1	other	0.106484																																																																											
3	from_this_person_to_poi	0.098848																																																																											

AdaBoostClassifier() with different features	Features importance																																																																																																
<pre>1 algo_list_ada=algo_list 2 clf_ada=AdaBoostClassifier() 3 pass_clf(algo_list_ada, features_list, output_list, data_df, clf_ada) 4</pre>	<table><tr><th></th><th>feature</th><th>importance</th></tr><tr><td>26</td><td>total_money</td><td>0.16</td></tr><tr><td>12</td><td>expenses</td><td>0.12</td></tr><tr><td>8</td><td>incentives_ratio</td><td>0.10</td></tr><tr><td>24</td><td>other</td><td>0.08</td></tr><tr><td>23</td><td>payment_f</td><td>0.08</td></tr><tr><td>19</td><td>exercised_stock_options</td><td>0.06</td></tr><tr><td>30</td><td>from_this_person_to_poi</td><td>0.04</td></tr><tr><td>29</td><td>total_stock_value</td><td>0.04</td></tr><tr><td>3</td><td>incentives</td><td>0.04</td></tr><tr><td>9</td><td>payment_2</td><td>0.04</td></tr><tr><td>22</td><td>total_payments</td><td>0.04</td></tr><tr><td>21</td><td>from_this_person_to_poi_ratio</td><td>0.04</td></tr><tr><td>1</td><td>restricted_stock</td><td>0.02</td></tr><tr><td>28</td><td>long_term_incentive</td><td>0.02</td></tr><tr><td>17</td><td>salary</td><td>0.02</td></tr><tr><td>0</td><td>to_messages</td><td>0.02</td></tr><tr><td>14</td><td>expenses_ratio</td><td>0.02</td></tr><tr><td>10</td><td>incentive_f</td><td>0.02</td></tr><tr><td>2</td><td>other_ratio</td><td>0.02</td></tr><tr><td>15</td><td>shared_receipt_with_poi</td><td>0.02</td></tr><tr><td>16</td><td>bonus_ratio</td><td>0.00</td></tr><tr><td>18</td><td>salary_ratio</td><td>0.00</td></tr><tr><td>13</td><td>from_poi_to_this_person</td><td>0.00</td></tr><tr><td>20</td><td>exercised_stock_options_ratio</td><td>0.00</td></tr><tr><td>11</td><td>payment_tt</td><td>0.00</td></tr><tr><td>7</td><td>from_messages</td><td>0.00</td></tr><tr><td>6</td><td>total_payments_ratio</td><td>0.00</td></tr><tr><td>25</td><td>emails_exc</td><td>0.00</td></tr><tr><td>5</td><td>total_stock_value_ratio</td><td>0.00</td></tr><tr><td>27</td><td>bonus</td><td>0.00</td></tr><tr><td>4</td><td>from_poi_to_this_person_ratio</td><td>0.00</td></tr></table>		feature	importance	26	total_money	0.16	12	expenses	0.12	8	incentives_ratio	0.10	24	other	0.08	23	payment_f	0.08	19	exercised_stock_options	0.06	30	from_this_person_to_poi	0.04	29	total_stock_value	0.04	3	incentives	0.04	9	payment_2	0.04	22	total_payments	0.04	21	from_this_person_to_poi_ratio	0.04	1	restricted_stock	0.02	28	long_term_incentive	0.02	17	salary	0.02	0	to_messages	0.02	14	expenses_ratio	0.02	10	incentive_f	0.02	2	other_ratio	0.02	15	shared_receipt_with_poi	0.02	16	bonus_ratio	0.00	18	salary_ratio	0.00	13	from_poi_to_this_person	0.00	20	exercised_stock_options_ratio	0.00	11	payment_tt	0.00	7	from_messages	0.00	6	total_payments_ratio	0.00	25	emails_exc	0.00	5	total_stock_value_ratio	0.00	27	bonus	0.00	4	from_poi_to_this_person_ratio	0.00
	feature	importance																																																																																															
26	total_money	0.16																																																																																															
12	expenses	0.12																																																																																															
8	incentives_ratio	0.10																																																																																															
24	other	0.08																																																																																															
23	payment_f	0.08																																																																																															
19	exercised_stock_options	0.06																																																																																															
30	from_this_person_to_poi	0.04																																																																																															
29	total_stock_value	0.04																																																																																															
3	incentives	0.04																																																																																															
9	payment_2	0.04																																																																																															
22	total_payments	0.04																																																																																															
21	from_this_person_to_poi_ratio	0.04																																																																																															
1	restricted_stock	0.02																																																																																															
28	long_term_incentive	0.02																																																																																															
17	salary	0.02																																																																																															
0	to_messages	0.02																																																																																															
14	expenses_ratio	0.02																																																																																															
10	incentive_f	0.02																																																																																															
2	other_ratio	0.02																																																																																															
15	shared_receipt_with_poi	0.02																																																																																															
16	bonus_ratio	0.00																																																																																															
18	salary_ratio	0.00																																																																																															
13	from_poi_to_this_person	0.00																																																																																															
20	exercised_stock_options_ratio	0.00																																																																																															
11	payment_tt	0.00																																																																																															
7	from_messages	0.00																																																																																															
6	total_payments_ratio	0.00																																																																																															
25	emails_exc	0.00																																																																																															
5	total_stock_value_ratio	0.00																																																																																															
27	bonus	0.00																																																																																															
4	from_poi_to_this_person_ratio	0.00																																																																																															
<p>CONFUSION MATRIX</p> <pre>[[37  3]  [ 1 31]]</pre>																																																																																																	
<table><tr><th>CLASSIFICATION</th><th>REPORT</th><th></th><th></th><th></th><th></th></tr><tr><td></td><td>precision</td><td>recall</td><td>f1-score</td><td>support</td><td></td></tr><tr><td>False</td><td>0.97</td><td>0.93</td><td>0.95</td><td>40</td><td></td></tr><tr><td>True</td><td>0.50</td><td>0.75</td><td>0.60</td><td>4</td><td></td></tr><tr><td>accuracy</td><td></td><td></td><td>0.91</td><td>44</td><td></td></tr><tr><td>macro avg</td><td>0.74</td><td>0.84</td><td>0.77</td><td>44</td><td></td></tr><tr><td>weighted avg</td><td>0.93</td><td>0.91</td><td>0.92</td><td>44</td><td></td></tr></table>	CLASSIFICATION	REPORT						precision	recall	f1-score	support		False	0.97	0.93	0.95	40		True	0.50	0.75	0.60	4		accuracy			0.91	44		macro avg	0.74	0.84	0.77	44		weighted avg	0.93	0.91	0.92	44																																																								
CLASSIFICATION	REPORT																																																																																																
	precision	recall	f1-score	support																																																																																													
False	0.97	0.93	0.95	40																																																																																													
True	0.50	0.75	0.60	4																																																																																													
accuracy			0.91	44																																																																																													
macro avg	0.74	0.84	0.77	44																																																																																													
weighted avg	0.93	0.91	0.92	44																																																																																													

<pre> algo_list_ada=['total_money','expenses','incentives_ratio','other','payment_f', 'exercised_stock_options','from_this_person_to_poi','total_stock_value', 'incentives','payment_2']  clf_ada=AdaBoostClassifier()  CONFUSION MATRIX [[38  2]  [ 2 21]]  CLASSIFICATION REPORT precision    recall  f1-score   support     False    0.95    0.95    0.95    40     True    0.50    0.50    0.50     4   accuracy    0.91    0.91    0.91    44  macro avg   0.72    0.72    0.73    44 weighted avg   0.91    0.91    0.91    44 </pre>	<pre> PRECISION 0.5  RECALL 0.5  feature importance 9      payment_2      0.16 0      total_money    0.14 1      expenses       0.12 5      exercised_stock_options 0.12 3      other          0.10 4      payment_f      0.10 6      from_this_person_to_poi 0.08 7      total_stock_value 0.08 8      incentives     0.06 2      incentives_ratio 0.04 </pre>
<pre> algo_list_ada2=['total_money','expenses','incentives_ratio', 'exercised_stock_options','other', 'from_this_person_to_poi' ]  clf_ada=AdaBoostClassifier()  CONFUSION MATRIX [[38  2]  [ 1 3]]  CLASSIFICATION REPORT precision    recall  f1-score   support     False    0.97    0.95    0.96    40     True    0.60    0.75    0.67     4   accuracy    0.93    0.93    0.93    44  macro avg   0.79    0.85    0.81    44 weighted avg   0.94    0.93    0.94    44 </pre>	<pre> PRECISION 0.6  RECALL 0.75  feature importance 4      other          0.24 1      expenses       0.20 3      exercised_stock_options 0.18 0      total_money    0.16 2      incentives_ratio 0.16 5      from_this_person_to_poi 0.06 </pre>

The features chosen will be determine for the algorithm used, in this case those are: 'poi', 'exercised\_stock\_options','other', 'expenses', 'total\_money', 'incentives\_ratio', 'payment\_f', 'from\_this\_person\_to\_poi', 'total\_stock\_value', "incentives", 'payment\_2'

Further details and code are available in PML\_4

3. What algorithm did you end up using? What other one(s) did you try? How did model performance differ between algorithms? [relevant rubric item: “pick an algorithm”]

As there are more non poi than poi, most probably we will get good ratios for prediction non poi than poi

The objective then is try to find an algorithm that in combination with some features provides a precision and recall are both at least 0.3

I will use as algorithm AdaBoostClassifier(base\_estimator=None, n\_estimators=500, learning\_rate=1.5, random\_state=42)

I have try several algorithms with original features, see table below. The best one are Logistic and RandomForest basic and with different parameters. There are Decission Tree classifier that identified 3 poi out of 4 but their error in identified non poi is too big to use them (>10 out of 40). They identifies several potential poi in the non poi population, we will comment on that later.

Algorithm with original features	Result
<pre> 1 # Comparing classifier with original features w/o big Nan features and after robusscaler 2 3 #train and test sets 4 f_list=algo_list0 5 set_train, set_test = train_test_split(data_df, test_size = 0.3, random_state=42) 6 features_train = set_train[f_list] 7 labels_train=set_train.poi 8 features_test = set_test[f_list ] 9 labels_test=set_test.poi 10 11 #Classifiers 12 clf1 = LogisticRegression(random_state=1) 13 clf2 = GaussianNB() 14 clf3 = AdaBoostClassifier() 15 clf4 = DecisionTreeClassifier() 16 clf5 = DecisionTreeClassifier(random_state=0) 17 clf6 = DecisionTreeClassifier(max_depth=None, min_samples_split=2,random_state=0) 18 clf7 = DecisionTreeClassifier(max_depth=None, min_samples_split=5,random_state=42) 19 clf8 = RandomForestClassifier() 20 clf9 = RandomForestClassifier(random_state=0) 21 clf10 = RandomForestClassifier(random_state=42) 22 clf11 = RandomForestClassifier(n_estimators=50, random_state=42) 23 clf12 = RandomForestClassifier(n_estimators=50, max_depth=None, min_samples_split=2, random_state=0) 24 clf13 = ExtraTreesClassifier(n_estimators=10, max_depth=None, min_samples_split=2, random_state=42) 25 26 #voting classifier 27 ecif = VotingClassifier( 28     estimators=[('Logistic Regression', clf1), ('naive Bayes', clf2), ('Ada', clf3), 29                 ('DecTre', clf4), ('DecTre0', clf5), ('DecTre20', clf6), ('DecTre542', clf7), 30                 ('RF', clf8), ('RF0', clf9), ('RF42', clf10), ('RF5042', clf11), ('RF5020', clf12), 31                 ('ET10242', clf13)], 32     voting='hard') 33 34 #classifier iteration and scores 35 for clf, label in zip([clf1, clf2, clf3,clf4,clf5,clf6,clf7,clf8,clf9,clf10,clf11, 36                       clf12,clf13,ecif], ['Logistic Regression', 'naive Bayes', 'Ada', 'RF', 'RF0', 37   'RF42', 'RF5042', 'RF5020', 'ET10242', 38   'Ensemble']): 39     scores = cross_val_score(clf, features_train, labels_train, scoring='accuracy', cv=5) 40     print("Accuracy: %0.2f (+/- %0.2f) [%s]" % (scores.mean(), scores.std(), label)) 41     fit = clf.fit(features_train, labels_train) 42     dt_pred = clf.predict(features_test) 43     labels_pred=clf.predict(features_test) 44 45     print("CONFUSION MATRIX", "\n", confusion_matrix(labels_test, labels_pred)) 46     dt_precision = precision_score(labels_test, dt_pred) 47     dt_recall = recall_score(labels_test, dt_pred) 48     print("PRECISION", dt_precision, "RECALL", dt_recall, "\n") </pre>	<p>Accuracy: 0.83 (+/- 0.06) [Logistic Regression]  CONFUSION MATRIX  [[35 1]  [3 1]]  PRECISION 0.5 RECALL 0.25</p> <p>Accuracy: 0.80 (+/- 0.04) [naive Bayes]  CONFUSION MATRIX  [[35 5]  [3 1]]  PRECISION 0.16666666666666666 RECALL 0.25</p> <p>Accuracy: 0.80 (+/- 0.08) [Ada]  CONFUSION MATRIX  [[35 5]  [2 2]]  PRECISION 0.2857142857142857 RECALL 0.5</p> <p>Accuracy: 0.76 (+/- 0.07) [DecTre]  CONFUSION MATRIX  [[29 11]  [1 3]]  PRECISION 0.21428571428571427 RECALL 0.75</p> <p>Accuracy: 0.72 (+/- 0.04) [DecTre0]  CONFUSION MATRIX  [[29 11]  [1 3]]  PRECISION 0.21428571428571427 RECALL 0.75</p> <p>Accuracy: 0.72 (+/- 0.04) [DecTre20]  CONFUSION MATRIX  [[29 11]  [1 3]]  PRECISION 0.21428571428571427 RECALL 0.75</p> <p>Accuracy: 0.74 (+/- 0.06) [DecTre542]  CONFUSION MATRIX  [[30 10]  [1 3]]  PRECISION 0.23076923076923078 RECALL 0.75</p> <p>Accuracy: 0.86 (+/- 0.05) [RF]  CONFUSION MATRIX  [[35 1]  [3 1]]  PRECISION 0.5 RECALL 0.25</p> <p>Accuracy: 0.85 (+/- 0.03) [RF0]  CONFUSION MATRIX  [[36 4]  [3 1]]  PRECISION 0.2 RECALL 0.25</p> <p>Accuracy: 0.86 (+/- 0.04) [RF42]  CONFUSION MATRIX  [[40 0]  [2 2]]  PRECISION 1.0 RECALL 0.5</p> <p>Accuracy: 0.86 (+/- 0.05) [RF5042]  CONFUSION MATRIX  [[36 4]  [3 1]]  PRECISION 0.2 RECALL 0.25</p> <p>Accuracy: 0.84 (+/- 0.03) [RF5020]  CONFUSION MATRIX  [[39 1]  [3 1]]  PRECISION 0.5 RECALL 0.25</p> <p>Accuracy: 0.85 (+/- 0.04) [ET10242]  CONFUSION MATRIX  [[36 2]  [3 1]]  PRECISION 0.3333333333333333 RECALL 0.25</p> <p>Accuracy: 0.84 (+/- 0.05) [Ensemble]  CONFUSION MATRIX  [[37 3]  [3 1]]  PRECISION 0.25 RECALL 0.25</p>

I have try several algorithms with original and created features, see table below. The best one are Ada and RandomForest basic and with different parameters. There are Decission Tree classifier that identified 2 poi out of 4 but their error in identified non poi is too big to use them (>5 out of 40).



Algorithm with original and created features	Result
<pre> 1 # 1. comparing classifier with 32 features after robustscaler 2 3 f_list_all=algo_list 4 set_train, set_test = train_test_split(data_df, test_size = 0.3, random_state=42) 5 features_train_a = set_train[f_list_all] 6 labels_train_a = set_train.poi 7 8 features_test_a = set_test[f_list_all] 9 labels_test_a = set_test.poi 10 11 clf1 = LogisticRegression(random_state=1) 12 clf2 = GaussianNB() 13 clf3 = AdaBoostClassifier() 14 clf4 = DecisionTreeClassifier() 15 clf5 = DecisionTreeClassifier(random_state=0) 16 clf6 = DecisionTreeClassifier(max_depth=None, min_samples_split=2, random_state=0) 17 clf7 = DecisionTreeClassifier(max_depth=None, min_samples_split=5, random_state=42) 18 clf8 = RandomForestClassifier() 19 clf9 = RandomForestClassifier(random_state=0) 20 clf10 = RandomForestClassifier(random_state=42) 21 clf11 = RandomForestClassifier(n_estimators=50, random_state=1) 22 clf12 = RandomForestClassifier(n_estimators=50, max_depth=None, min_samples_split=2, random_state=0) 23 clf13 = ExtraTreesClassifier(n_estimators=10, max_depth=None, min_samples_split=2, random_state=42) 24 25 eclf = VotingClassifier( 26     estimators=[('Logistic Regression', clf1), ('naive Bayes', clf2), ('Ada', clf3), 27                 ('DecTre', clf4), ('DecTre0', clf5), ('DecTre20', clf6), ('DecTre542', clf7), 28                 ('RF', clf8), ('RF0', clf9), ('RF42', clf10), ('RF501', clf11), ('RF5020', clf12), 29                 ('ET10242', clf13)], 30     voting='hard') 31 32 for clf, label in zip([clf1, clf2, clf3, clf4, clf5, clf6, clf7, clf8, clf9, clf10, clf11, 33                       clf12, clf13, eclf], ['Logistic Regression', 'naive Bayes', 'Ada', 34   'DecTre', 'DecTre0', 'DecTre20', 'DecTre542', 35   'RF', 'RF0', 'RF42', 'RF501', 'RF5020', 'ET10242', 36   'Ensemble']): 37     scores = cross_val_score(clf, features_train_a, labels_train_a, scoring='accuracy', cv=5) 38     print("Accuracy: %0.2f (+/- %0.2f) [%s]" % (scores.mean(), scores.std(), label)) 39     fit = clf.fit(features_train_a, labels_train_a) 40     dt_pred_a = clf.predict(features_test_a) 41     labels_pred_a = clf.predict(features_test_a) 42 43     print("CONFUSION MATRIX", "\n", confusion_matrix(labels_test, labels_pred)) 44     dt_precision = precision_score(labels_test, dt_pred) 45     dt_recall = recall_score(labels_test, dt_pred) 46     print("PRECISION", dt_precision, "RECALL", dt_recall, "\n") </pre>	<p>Accuracy: 0.82 (+/- 0.07) [Logistic Regression] CONFUSION MATRIX [35 5] [2 2] PRECISION 0.2857142857142857 RECALL 0.5</p> <p>Accuracy: 0.78 (+/- 0.07) [naive Bayes] CONFUSION MATRIX [35 5] [3 1] PRECISION 0.16666666666666666 RECALL 0.25</p> <p>Accuracy: 0.85 (+/- 0.03) [Ada] CONFUSION MATRIX [37 3] [1 3] PRECISION 0.5 RECALL 0.75</p> <p>Accuracy: 0.80 (+/- 0.07) [DecTre] CONFUSION MATRIX [35 5] [3 1] PRECISION 0.16666666666666666 RECALL 0.25</p> <p>Accuracy: 0.81 (+/- 0.05) [DecTre0] CONFUSION MATRIX [34 6] [4 0] PRECISION 0.0 RECALL 0.0</p> <p>Accuracy: 0.81 (+/- 0.05) [DecTre20] CONFUSION MATRIX [34 6] [4 0] PRECISION 0.0 RECALL 0.0</p> <p>Accuracy: 0.81 (+/- 0.04) [DecTre542] CONFUSION MATRIX [33 7] [2 2] PRECISION 0.2222222222222222 RECALL 0.5</p> <p>Accuracy: 0.84 (+/- 0.04) [RF] CONFUSION MATRIX [39 1] [4 0] PRECISION 0.0 RECALL 0.0</p> <p>Accuracy: 0.84 (+/- 0.03) [RF0] CONFUSION MATRIX [38 2] [4 0] PRECISION 0.0 RECALL 0.0</p> <p>Accuracy: 0.83 (+/- 0.03) [RF42] CONFUSION MATRIX [38 2] [4 0] PRECISION 0.0 RECALL 0.0</p> <p>Accuracy: 0.86 (+/- 0.04) [RF501] CONFUSION MATRIX [39 1] [3 1] PRECISION 0.5 RECALL 0.25</p> <p>Accuracy: 0.86 (+/- 0.03) [RF5020] CONFUSION MATRIX [39 1] [3 1] PRECISION 0.5 RECALL 0.25</p> <p>Accuracy: 0.86 (+/- 0.03) [ET10242] CONFUSION MATRIX [38 2] [2 2] PRECISION 0.5 RECALL 0.5</p> <p>Accuracy: 0.85 (+/- 0.04) [Ensemble] CONFUSION MATRIX [37 3] [3 1] PRECISION 0.25 RECALL 0.25</p>

Some algorithm perform well in identifying non poi that means they do not perform well identifying poi. The same algorithm perform differently depending on the features we pass to it.

Logistic regression with original features ( 'salary', 'bonus', 'long\_term\_incentive', 'other', 'expenses', 'total\_payments', 'exercised\_stock\_options', 'restricted\_stock', 'total\_stock\_value', 'from\_messages', 'to\_messages', 'from\_poi\_to\_this\_person', 'from\_this\_person\_to\_poi', 'shared\_receipt\_with\_poi') identifies 39 non poi but only 1 poi. When we pass to it the original and the created features, it identifies only 35 non poi but 2 poi, it means it define what is a poi better but it means more non poi fall in the poi characterization

Decission tree with original features identify 3 poi out of 4 but identify as poi 11 non poi, what means the threshold to become poi is allowing a lot of non poi to fall in that category. The classifier is weak in non poi identification. When we pass all the features to Decission tree, it increase the accuracy on non poi (only 5 non poi identified as poi) but reduce the identification of poi (1 out of 4) it means increase the error on poi.

We need to define which features to pass at same time we fine tune the algorithm that is an iterative activity supported by other algorithm and combination of machine intelligence and human intuition.

RandomForest with original features and random = 42 identifies 40 non poi out of 40 and 2 poi out of 4. The accuracy us 0.86, the precision is 1 and the recall 0.5

RandomForest with original and created features identifies 38 non poi out of 40 and 0 poi out of 4. The accuracy us 0.83, the precision is 0 and the recall 0



AdaBoost with original features identifies 35 non poi out of 40 and 2 poi out of 4. The accuracy is 0.80, the precision is 0.28 and the recall 0.5

AdaBoost with original and created features identifies 37 non poi out of 40 and 3 poi out of 4. The accuracy is 0.85, the precision is 0.5 and the recall 0.75

I will use a combination of original and created features while fine tune the two chosen algorithm: RandomForest and AdaBoost

Further details and code are available in PML\_4

- What does it mean to tune the parameters of an algorithm, and what can happen if you don't do this well? How did you tune the parameters of your particular algorithm? What parameters did you tune?

The parameters of an algorithm are defined by default but it could be adjusted to get the algorithm to improve the result and provide a more appropriated answer to the question.

This project aim is to split the people involved in Enron fraud in two cluster: poi and non-poi. The different algorithm provides different answer depending on their parameters and on the features we use. We need to choose the one that answer better to our question.

The tune of the parameters allow the analyst to improve the performance and to move the sensibility of the algorithm in order to minimize false negative and false positive.

My intention with this project is to reduce false negatives, that will increase false positive, that means that I want my algorithm to find as much as poi as possible and to detect over the non poi population which one could be considered poi. That is like as after reading some documentation, I have arrive to the conclusion that some of actors in Enron fraud have collaborate with the investigation and in exchange the got the consideration of non poi.

We can manually optimize the parameters or use some optimization algorithm. I have use GridSearchCV to fine tune random forest.

We can use RandomizedSearchCV to define the range of the values for the hyperparametres or we can define the range to check over GridSearch by our own criteria and manually fine-tune.

Random forest I checked for 'n\_estimators':[200, 400, 600], 'max\_depth':[10, 20, 30, 40, 50, 60], 'min\_samples\_split': [2, 5, 10], 'min\_samples\_leaf': [1, 2, 4], 'bootstrap': [True, False] and 'random\_state':[42] but the result was not as good as I wanted then I try again with { 'n\_estimators': [10,200,500], 'max\_depth':[7,15,24,200,500], 'max\_features':[2,5], 'random\_state':[42] } and the result are in the table below

GridSearchCV on Random Forest	Result																																																																																										
<pre>algo_list_rf=['exercised_stock_options','other','bonus','from_this_person_to_poi','total_payments' ]  set_train, set_test = train_test_split(data_df, test_size = 0.3, random_state=42) features_train = set_train[algo_list_rf] labels_train=set_train.poi  features_test = set_test[algo_list_rf] labels_test=set_test.poi  from sklearn.ensemble import RandomForestClassifier clf_rf=RandomForestClassifier(random_state=42)  param_grid={     'n_estimators': [10,200,500],     'max_depth':[7,15,24,200,500],     'max_features':[2,5],     'random_state':[42] }  f1=make_scorer(f1_score, average='macro') CV_clf_rf=GridSearchCV(estimator=clf_rf,param_grid=param_grid )  CV_clf_rf.fit(features_train, labels_train) CV_clf_rf.best_params_</pre>	<pre>(,max"qetcp,1 1' ,max"gewetces,1 5' ,u"eactrewetces,1 500' ,eetqetw"etwec,1 45)</pre>																																																																																										
<pre># algorithm with result of GridSearch and random state=42  algo_list_rf=['exercised_stock_options','other','bonus','from_this_person_to_poi','total_payments' ]  clf_rf1=RandomForestClassifier(n_estimators=200, max_depth=7, max_features = 2, random_state=42)  features_list_rf=algo_list_rf.copy() features_list_rf.insert(0,'poi')  output_list_rf=features_list_rf.copy() output_list_rf.insert(1,'poi_pred')  pass_clf(algo_list_rf, features_list_rf, output_list_rf, data_df, clf_rf1)</pre>	<p>CONFUSION MATRIX</p> <pre>[[38  2]  [ 2  2]]</pre> <p>CLASSIFICATION REPORT</p> <table><thead><tr><th></th><th>precision</th><th>recall</th><th>f1-score</th><th>support</th></tr></thead><tbody><tr><td>False</td><td>0.95</td><td>0.95</td><td>0.95</td><td>40</td></tr><tr><td>True</td><td>0.50</td><td>0.50</td><td>0.50</td><td>4</td></tr><tr><td>accuracy</td><td></td><td></td><td>0.91</td><td>44</td></tr><tr><td>macro avg</td><td>0.72</td><td>0.72</td><td>0.73</td><td>44</td></tr><tr><td>weighted avg</td><td>0.91</td><td>0.91</td><td>0.91</td><td>44</td></tr></tbody></table>		precision	recall	f1-score	support	False	0.95	0.95	0.95	40	True	0.50	0.50	0.50	4	accuracy			0.91	44	macro avg	0.72	0.72	0.73	44	weighted avg	0.91	0.91	0.91	44																																																												
	precision	recall	f1-score	support																																																																																							
False	0.95	0.95	0.95	40																																																																																							
True	0.50	0.50	0.50	4																																																																																							
accuracy			0.91	44																																																																																							
macro avg	0.72	0.72	0.73	44																																																																																							
weighted avg	0.91	0.91	0.91	44																																																																																							
<table><thead><tr><th></th><th>poi</th><th>poi_pred</th><th>exercised_stock_options</th><th>other</th><th>bonus</th><th>from_this_person_to_poi</th><th>total_payments</th></tr></thead><tbody><tr><td>KOENIG MARK E</td><td>True</td><td>False</td><td>0.04</td><td>1.01</td><td>0.50</td><td>1.07</td><td>0.37</td></tr><tr><td>LAY KENNETH L</td><td>True</td><td>True</td><td>20.04</td><td>69.72</td><td>8.38</td><td>1.14</td><td>57.10</td></tr><tr><td>HANNON KEVIN P</td><td>True</td><td>True</td><td>2.93</td><td>0.07</td><td>1.50</td><td>1.50</td><td>-0.35</td></tr><tr><td>COLWELL WESLEY</td><td>True</td><td>False</td><td>-0.36</td><td>0.88</td><td>1.12</td><td>0.79</td><td>0.32</td></tr><tr><td>LAY KENNETH L</td><td>True</td><td>True</td><td>20.04</td><td>69.72</td><td>8.38</td><td>1.14</td><td>57.10</td></tr><tr><td>HANNON KEVIN P</td><td>True</td><td>True</td><td>2.93</td><td>0.07</td><td>1.50</td><td>1.50</td><td>-0.35</td></tr><tr><td>HICKERSON GARY J</td><td>False</td><td>True</td><td>-0.36</td><td>0.01</td><td>1.75</td><td>0.07</td><td>0.65</td></tr><tr><td>LAVORATO JOHN J</td><td>False</td><td>True</td><td>2.11</td><td>0.00</td><td>9.62</td><td>29.36</td><td>5.29</td></tr></tbody></table>		poi	poi_pred	exercised_stock_options	other	bonus	from_this_person_to_poi	total_payments	KOENIG MARK E	True	False	0.04	1.01	0.50	1.07	0.37	LAY KENNETH L	True	True	20.04	69.72	8.38	1.14	57.10	HANNON KEVIN P	True	True	2.93	0.07	1.50	1.50	-0.35	COLWELL WESLEY	True	False	-0.36	0.88	1.12	0.79	0.32	LAY KENNETH L	True	True	20.04	69.72	8.38	1.14	57.10	HANNON KEVIN P	True	True	2.93	0.07	1.50	1.50	-0.35	HICKERSON GARY J	False	True	-0.36	0.01	1.75	0.07	0.65	LAVORATO JOHN J	False	True	2.11	0.00	9.62	29.36	5.29	<p>PRECISION</p> <p>0.5</p> <p>RECALL</p> <p>0.5</p> <table><thead><tr><th></th><th>feature</th><th>importance</th></tr></thead><tbody><tr><td>0</td><td>exercised_stock_options</td><td>0.254282</td></tr><tr><td>2</td><td>bonus</td><td>0.228429</td></tr><tr><td>1</td><td>other</td><td>0.204395</td></tr><tr><td>4</td><td>total_payments</td><td>0.178578</td></tr><tr><td>3</td><td>from_this_person_to_poi</td><td>0.134316</td></tr></tbody></table>		feature	importance	0	exercised_stock_options	0.254282	2	bonus	0.228429	1	other	0.204395	4	total_payments	0.178578	3	from_this_person_to_poi	0.134316
	poi	poi_pred	exercised_stock_options	other	bonus	from_this_person_to_poi	total_payments																																																																																				
KOENIG MARK E	True	False	0.04	1.01	0.50	1.07	0.37																																																																																				
LAY KENNETH L	True	True	20.04	69.72	8.38	1.14	57.10																																																																																				
HANNON KEVIN P	True	True	2.93	0.07	1.50	1.50	-0.35																																																																																				
COLWELL WESLEY	True	False	-0.36	0.88	1.12	0.79	0.32																																																																																				
LAY KENNETH L	True	True	20.04	69.72	8.38	1.14	57.10																																																																																				
HANNON KEVIN P	True	True	2.93	0.07	1.50	1.50	-0.35																																																																																				
HICKERSON GARY J	False	True	-0.36	0.01	1.75	0.07	0.65																																																																																				
LAVORATO JOHN J	False	True	2.11	0.00	9.62	29.36	5.29																																																																																				
	feature	importance																																																																																									
0	exercised_stock_options	0.254282																																																																																									
2	bonus	0.228429																																																																																									
1	other	0.204395																																																																																									
4	total_payments	0.178578																																																																																									
3	from_this_person_to_poi	0.134316																																																																																									

I use a similar approach to ADABOOST but with the hyperparametres related to ADABOOST ('n\_estimators':[7,15,24,200,500,900], 'learning\_rate':[1, 1.5, 2, 2.5, 3,], 'random\_state':[42]) with a result as good as for random forest as show in table below

GridSearchCV on AdaBoost

```
algo_list_ada=['total_money','expenses','incentives_ratio',
               'exercised_stock_options', 'other', 'from_this_person_to_poi' ]

set_train, set_test = train_test_split(data_df, test_size = 0.3, random_state=42)
features_train = set_train[algo_list_ada]
labels_train=set_train.poi

features_test = set_test[algo_list_ada]
labels_test=set_test.poi

clf_ada=AdaBoostClassifier(random_state=42)

param_grid={'n_estimators':[7,15,24,200,500,900],
            'learning_rate':[1, 1.5, 2, 2.5, 3,],
            'random_state':[42]
            }

f1=make_scorer(f1_score, average='macro')
CV_clf_rf=GridSearchCV(estimator=clf_ada,param_grid=param_grid )

CV_clf_rf.fit(features_train, labels_train)
CV_clf_rf.best_params_

algo_list_ada=['total_money','expenses','incentives_ratio',
               'exercised_stock_options', 'other', 'from_this_person_to_poi'
               ]
clf_ada=AdaBoostClassifier(base_estimator=None, n_estimators=15, learning_rate=1.5,
                           random_state=42)

features_list_ada=algo_list_ada.copy()
features_list_ada.insert(0,'poi')

output_list_ada=features_list_ada.copy()
output_list_ada.insert(1,'poi_pred')

pass_clf(algo_list_ada,features_list_ada, output_list_ada, data_df, clf_ada)
```

Result

```
{'learning_rate': 1.5, 'n_estimators': 15, 'random_state': 42}
```

CONFUSION MATRIX

```
[[36  4]
 [ 1  3]]
```

CLASSIFICATION REPORT

	precision	recall	f1-score	support
False	0.97	0.90	0.94	40
True	0.43	0.75	0.55	4
accuracy			0.89	44
macro avg	0.70	0.82	0.74	44
weighted avg	0.92	0.89	0.90	44

	poi_pred	total_money	expenses	incentives_ratio	exercised_stock_options	other	from_this_person_to_poi	
KOENIG MARK E	True	True	0.31	1.94	0.88	0.04	1.01	1.07
LAY KENNETH L	True	True	32.00	1.44	0.53	20.04	69.72	1.14
HANNON KEVIN P	True	False	0.98	0.22	2.40	2.93	0.07	1.50
COLWELL WESLEY	True	True	0.03	-0.10	0.00	-0.36	0.68	0.79
KOENIG MARKE	True	True	0.31	1.94	0.88	0.04	1.01	1.07
LAY KENNETH L	True	True	32.00	1.44	0.53	20.04	69.72	1.14
KISHKILL JOSEPH G	False	True	-0.07	1.74	0.00	-0.35	3.13	0.00
PIPER GREGORY F	False	True	0.12	0.39	0.00	0.16	-0.00	3.43
HICKERSON GARY J	False	True	0.10	1.42	0.00	-0.36	0.01	0.07
COLWELL WESLEY	True	True	0.03	-0.10	0.00	-0.36	0.68	0.79
MCQUELLAN GEORGE	False	True	0.04	3.81	0.00	-0.06	0.34	0.00

PRECISION

```
0.42857142857142855
```

RECALL

```
0.75
```

	feature	importance
0	total_money	0.200000
3	exercised_stock_options	0.200000
4	other	0.200000
5	from_this_person_to_poi	0.200000
2	incentives_ratio	0.139333
1	expenses	0.066667

Further details and code are available in PML\_4

5. What is validation, and what's a classic mistake you can make if you do it wrong?  
How did you validate your analysis?

I have use train\_test\_split validation strategy along the project.

Since the beginning the data has been split in train and test set to be able to estimate the performance of the model in an independent data set and check overfitting.

The point is that if we train the model with all the dataset when we try to get what is the performance of the model, we do not have data to pass to the model.

I have used too StratifiedShuffleSplit as provided in tester.py

train\_test\_split

```
set_train, set_test = train_test_split(data_df, test_size = 0.3, random_state=42)
features_train = set_train[algo_list]
labels_train=set_train.poi

features_test = set_test[algo_list ]
labels_test=set_test.poi
```

StratifiedShuffleSplit

```
def test_train(dataset, feature_list, folds=1000):
    data = featureFormat(dataset, feature_list, sort_keys=True)
    labels, features = targetFeatureSplit(data)
    # cv = StratifiedShuffleSplit(labels, folds, random_state=42)
    cv = StratifiedShuffleSplit(n_splits=folds, random_state=42)
    true_negatives = 0
    false_negatives = 0
    true_positives = 0
    false_positives = 0
    #for train_idx, test_idx in cv:
    for train_idx, test_idx in cv.split (features, labels):
        features_train = []
        features_test = []
        labels_train = []
        labels_test = []
        for ii in train_idx:
            features_train.append(features[ii])
            labels_train.append(labels[ii])
        for jj in test_idx:
            features_test.append(features[jj])
            labels_test.append(labels[jj])
    return features_train, features_test, labels_train, labels_test
```

6. Give at least 2 evaluation metrics and your average performance for each of them. Explain an interpretation of your metrics that says something human-understandable about your algorithm's performance.

I train the algorithm with train set and then pass the test features to algorithm to predict and then calculate the different performance measures: accuracy (good answer vs all answer), precision (true positives vs all identified as positives), recall (true positives vs all positives)

The request from the project is to get a minimum of 0.3 recall and 0.3 precision.

The population of 144 entries, is split at test size 0.3, what means a test size of 40 entries where 4 are poi and 40 non poi. The best way to train and test this model with this low level of data is to iterate the train and test set using a cross validation object as could be StratifiedShuffleSplit

I have pass different classifier to tester and chose the one with better result to submit as final classifier

Classifier, features, metrics results and explanations

RandomForest finetune

```
clf_rf2=RandomForestClassifier(n_estimators=200, max_depth=7, max_features = 2, random_state=42)
features_rfd=['poi', 'bonus', 'exercised_stock_options', 'total_stock_value', 'expenses',
              'total_payments', 'incentives_ratio' ]

# Define pipeline for scaler and classifier
clf_trial=clf_rf2
pipe_clf= Pipeline([['rob_sc', RobustScaler() ], ['clf', clf_trial]])
clf_rf2p=pipe_clf
# Test the classifier
print('RandomForestClassifier(finetune)')
test_classifier(clf_rf2p, my_dataset, features_rfd, folds=1000)

RandomForestClassifier(finetune)
Pipeline(memory=None,
       steps=[('rob_sc',
               RobustScaler(copy=True, quantile_range=(25.0, 75.0),
                             with_centering=True, with_scaling=True)),
              ('clf',
               RandomForestClassifier(bootstrap=True, class_weight=None,
                                       criterion='gini', max_depth=7,
                                       max_features=2, max_leaf_nodes=None,
                                       min_impurity_decrease=0.0,
                                       min_impurity_split=None,
                                       min_samples_leaf=1, min_samples_split=2,
                                       min_weight_fraction_leaf=0.0,
                                       n_estimators=200, n_jobs=None,
                                       oob_score=False, random_state=42,
                                       verbose=0, warm_start=False))],
       verbose=False)
Accuracy: 0.87693      Precision: 0.57505      Recall: 0.29500 F1: 0.38995      F2: 0.32683
Total predictions: 15000      True positives: 590      False positives: 436      False negatives: 1410      True negatives: 12564
```

Those metrics are the average metrics of all the iterations of the classifier over the dataset

Accuracy: 87% of the answer are good (true / total). That means there are 13% probability to mistake when indicate someone is poi or non poi

Regarding the quality for the poi identification we have precision and recall.

Precision: 57% of the people indicate as poi are really poi (true + / true and false +). That means that if someone is identified as poi, there are 43% of probability this person is non poi.

Recall: 29% of poi are identified as poi (true + / true+ plus false -). That means that if someone is poi, there are 29% of probability the algorithm identified this person as poi.

Total predictions: 15000

True positives: 590

False positives: 436 (algorithm identified them as poi but they are declared non-poi)

False negatives: 1410 (algorithm identified as non-poi but they are declared as poi)

True negatives: 12564

## AdaBoost (base estimator RandomForest and finetune)

```
clf_rf2=RandomForestClassifier(n_estimators=200, max_depth=7, max_features = 2, random_state=42)
clf_ada_rf2=AdaBoostClassifier(base_estimator=clf_rf2, n_estimators=500, learning_rate=1.5, random_state=42)
features_rfd=['poi', 'bonus','exercised_stock_options', 'total_stock_value', 'expenses' ,
              'total_payments','incentives_ratio' ]

# Define pipeline for scaler and classifier
clf_trial=clf_ada_rf2
pipe_clf= Pipeline([['rob_sc',RobustScaler() ],['clf', clf_trial]])
clf_ada_rf2p=pipe_clf
# Test the classifier
print('AdaBoostClassifier(base estimator RandomForest and finetune)')
test_classifier(clf_ada_rf2p, my_dataset, features_rfd, folds=1000)

AdaBoostClassifier(base estimator RandomForest and finetune)
Pipeline(memory=None,
         steps=[('rob_sc',
                 RobustScaler(copy=True, quantile_range=(25.0, 75.0),
                               with_centering=True, with_scaling=True)),
                ('clf',
                 AdaBoostClassifier(algorithm='SAMME.R',
                                     base_estimator=RandomForestClassifier(bootstrap=True,
                                     class_weight=None,
                                     criterion='gini',
                                     max_depth=7,
                                     max_features=2,
                                     max_leaf_nodes=None,
                                     min_impurity_decrease=0.0,
                                     min_impurity_split=None,
                                     min_samples_leaf=1,
                                     min_samples_split=2,
                                     min_weight_fraction_leaf=0.0,
                                     n_estimators=200,
                                     n_jobs=None,
                                     oob_score=False,
                                     random_state=42,
                                     verbose=0,
                                     warm_start=False),
                                     learning_rate=1.5, n_estimators=500,
                                     random_state=42)]],
         verbose=False)
Accuracy: 0.87767      Precision: 0.58144      Recall: 0.29450 F1: 0.39097      F2: 0.32675
Total predictions: 15000      True positives: 589      False positives: 424      False negatives: 1411      True
negatives: 12576
```

Those metrics are the average metrics of all the iterations of the classifier over the dataset

Accuracy: 88% of the answer are good (true / total). That means there are 12% probability to mistake when indicate someone is poi or non poi

Regarding the quality for the poi identification we have precision and recall.

Precision: 58% of the people indicate as poi are really poi (true + / true and false +). That means that if someone is identified as poi, there are 42% of probability this person is non poi.

Recall: 29% of poi are identified as poi (true + / true+ plus false -). That means that if someone is poi, there are 29% of probability the algorithm identified this person as poi.

Total predictions: 15000

True positives: 589

False positives: 424 (algorithm identified them as poi but they are declared non-poi)

False negatives: 1411 (algorithm identified as non-poi but they are declared as poi)

True negatives: 12576

### AdaBoost (finetune)

```
clf_ada2=AdaBoostClassifier(base_estimator=None, n_estimators=500, learning_rate=1.5, random_state=42)
features_adah=['poi', 'exercised_stock_options', 'other', 'expenses', 'total_money', 'incentives_ratio',
               'from_this_person_to_poi',
               'total_stock_value', 'payment_f', 'incentives', 'payment_2']

# Define pipeline for scaler and classifier
clf_trial=clf_ada2
pipe_clf= Pipeline([['rob_sc', RobustScaler() ], ['clf', clf_trial]])
clf_ada2p=pipe_clf
# Test the classifier
print('AdaBoostClassifier(finetune)')
test_classifier(clf_ada2p, my_dataset, features_adah, folds=1000)

AdaBoostClassifier(finetune)
Pipeline(memory=None,
       steps=[('rob_sc',
               RobustScaler(copy=True, quantile_range=(25.0, 75.0),
                             with_centering=True, with_scaling=True)),
              ('clf',
               AdaBoostClassifier(algorithm='SAMME.R', base_estimator=None,
                                   learning_rate=1.5, n_estimators=500,
                                   random_state=42))],
       verbose=False)
Accuracy: 0.89253      Precision: 0.65696      Recall: 0.40600 F1: 0.50185      F2: 0.43958
Total predictions: 15000      True positives: 812      False positives: 424      False negatives: 1188      True
negatives: 12576
```

Those metrics are the average metrics of all the iterations of the classifier over the dataset

Accuracy: 89% of the answer are good (true / total). That means there are 11% probability to mistake when indicate someone is poi or non poi

Regarding the quality for the poi identification we have precision and recall.

Precision: 66% of the people indicate as poi are really poi (true + / true and false +). That means that if someone is identified as poi, there are 34% of probability this person is non poi.

Recall: 40% of poi are identified as poi (true + / true+ plus false -). That means that if someone is poi, there are 40% of probability the algorithm identified this person as poi.

Total predictions: 15000

True positives: 812

False positives: 424 (algorithm identified them as poi but they are declared non-poi)

False negatives: 1188 (algorithm identified as non-poi but they are declared as poi)

True negatives: 12576

This combination of classifier with hyper parameters and features is the one chosen to be used as final classifier. It covers the requirement: When tester.py is used to evaluate performance, precision and recall are both at least 0.3.

My conclusion is that the machine learning could identify poi base on the information we give but the reason why some people has not been considered poi is more complex than the information we have. There were a lot of political implication and on top some people probably give information to the investigation on exchange of not being considered as poi.

But that is my thought after reading the press and passing the data over machine learning getting once and again this person as poi being considered not poi

Along the project I have used different techniques to find the best combination of features, classifier and hyper parameters but the final finetune has been done manually in a trial-error process.

It is important to keep in mind that algorithm help a lot to understand and predict but we need to ensure data quality and common sense in the information we pass to the algorithm. Sometimes we are lucky and the algorithm is able to provide some indications about the weakness in the data we pass.

This is by far the most interesting project of the training as it requires to use mainly all the knowledge develop in the previous projects. It is a project that confront you with real data and real situation where multiples ways of solving the question could be applied. You need to keep in mind the objective of the project as the ramifications could be very long and attractive. I have really enjoy doing it.

I fell ready know to use machine learning technique in my daily work convinced that there are still a lot to learn and to try but that is very funny.