1. **Summarize for us the goal of this project and how machine learning is useful in trying to accomplish it. As part of your answer, give some background on the dataset and how it can be used to answer the project question. Were there any outliers in the data when you got it, and how did you handle those?**

Enron was one of the world's largest energy and service company in Texas, US, before its bankruptcy on 2001. It was revealed of institutionalized and systematic accounting fraud. During the investigation by FERC, the financial and email information of near 150 Enron employees were made open to public.

The goal of this project is to identify person of interest (POI) in the Enron fraud case. The dataset contains 146 records with 18 POIs and the rest non-POIS. For each record, it contains 14 financial features, 6 email features, and 1 POI label. There are about 40% of the feature values given with 'NaN' and will be replaced by 0.0. Machining learning will be used to train models to predict whether a person is POI based on the financial and email information.

Before building the model, the variables are examined for outliers. 'Total' is an outlier seen on the scatter plot of salary and bonus. It is removed because it is only a sum of other records. For each variable, potential outliers (value higher than 98 percentile) are checked with "enron61702insiderpay.pdf". It is found the data for BELFER ROBERT and BHATNAGAR SANJAY are under wrong variable names, and give outliers in "restricted_stock_deferred". They are corrected according to the origin data sheet. Besides, all NaN are converted to 0 and data point with all features 0 are omitted, and it removes LOCKHART EUGENE E.

2. **What features did you end up using in your POI identifier, and what selection process did you use to pick them? Did you have to do any scaling? Why or why not? As part of the assignment, you should attempt to engineer your own feature that does not come ready-made in the dataset -- explain what feature you tried to make, and the rationale behind it. In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importance of the features that you use, and if you used an automated feature selection function like SelectKBest, please report the feature scores and reasons for your choice of parameter values.**

I add new features "fraction_from_poi" and "fraction_to_poi", which are the fraction of messages to/from that person that are from/to a POI. It's expected that POIs might send emails to each other at higher rate. Another feature created is "fraction_salary", the fraction of salary to total payments. It's possible that POI have different proportions of the payments.

Then decision tree algorithm is used to show the feature importance. The importance of the used features:

1. exercised_stock_options 0.31
2. expenses 0.26
3. other 0.23
4. shared_receipt_with_poi 0.20

Iterative process is used for feature selection. First, decision tree is performed with all the features. Then features with importance less than 0.1 or the lowest are removed. And decision tree is performed with the rest features. For each step, precision and recall are calculated. The process ends when the highest score is reached. The new features are not used in the end. The performance with these features has better precision but lower recall, and overall F1 is lower.

	Precision	Recall	F1
With fraction_from/to_poi	0.52	0.34	0.41
With fraction_from/to_poi , fraction_salary	0.46	0.39	0.42
Without any new features	0.43	0.43	0.43

No scaling is needed at this stage, since decision tree algorithm doesn't need scaling.

3. What algorithm did you end up using? What other one(s) did you try? How did model performance differ between algorithms?

Finally, I decide to use the Decision Tree algorithm. I also tried Gaussian Naive Bayes, Random Forest, and Support Vector Machine. The precision, recall and f1-score for each algorithm are summarized in the table below.

Algorithm	Precision	Recall	F1-score
Decision Tree	0.43	0.43	0.43
Gaussian NB	0.42	0.23	0.30
Random Forest	0.55	0.16	0.25
SVM*	0.85	0.20	0.33

*Scaling is performed for SVM.

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?

Most of the machine learning algorithms are parameterized. The parameters affect the final performance of the model in precision, recall, f1-score. Not tuning the parameters well may lead the algorithms not reaching its best performance or over-fitting.

I use the decision tree algorithm. I tried different combination of values for the decision tree parameters "splitter", "max_features", "min_samples_split", "criterion",

and “min_sample_leaf”, by passing these lists of parameters to the GridSearchCV. The cv use the StratifiedShuffleSplit method. I also compared the results with and without PCA. And the optimized value based on f1-score are returned and used in the model.

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

Validation is a process in machine learning to estimate the performance of the model on new dataset. Validation can help prevent from over fitting, where the model perform well on the training dataset but poor with new data. Therefore, data are splitted into training set and testing set. The model is trained with the training set and evaluated on the test set. A classic mistake is not to split the data randomly, and the training set mostly contains data of one category while the testing set contains another category.

In this project, the data is splitted using the “StratifiedShuffleSplit”, with 1000 folds. The method returns stratified randomized folds, with each fold preserving the percentage of samples for each class.

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.**

Two of the most important evaluation metrics are precision and recall. For the decision tree model used in this project, the precision is 0.43 and recall is 0.43. The precision means that 43% of the people who are predicted as POI by the model are real POI. The recall value means that 43% of all the real POI are predicted by the model as POI.

Reference

1. https://en.wikipedia.org/wiki/Enron_scandal
2. https://www.cs.cmu.edu/~./enron/
3. http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html#sklearn.tree.DecisionTreeClassifier
4. http://scikit-learn.org/stable/modules/svm.html
5. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
6. http://scikit-learn.org/stable/modules/generated/sklearn.pipeline.Pipeline.html
7. http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedShuffleSp

lit.html
8. http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html
9. http://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html