Udacity Data Analyst Nanodegree
Project 5: Identify Fraud from Enron Email

1) The goal of this project is to identify persons involved in fraud in the Enron scandal (Persons of Interest). The dataset consists of a list of persons involved with Enron and their financial attributes. It has variables such as salary, bonus, stock value, etc. These can be used with machine learning techniques to identify patterns of potential persons of interest.

   The dataset contains a few outliers; there were a few data points, which had some values for total payments much higher than other nodes. A linear regression was used in an attempt to estimate data points with total payments, which differed significantly from the dataset, 7% of the records with the highest error were removed.

   Also, for a few persons, all variables had no values. These included: Cline Kenneth W, Piro Jim, Lockhart Eugene E, Hayslett Roderick J. Without these, there were 146 records in the dataset, by removing 7% of outliers 133 records were kept for further analysis.

2) The dataset contains several features. Not all of them were used for the machine-learning algorithm. The features were selected based on a Random Forest Classifier. The features with an importance higher than 8% were kept for further analysis, these included and their importance were: total payments (11.9%), exercised stock options (10.5%), bonus (11.0%) and deferred income (8.2%).

   No scaling was done with those features, but instead they were a Principal Component Analysis was used to create three new features.

3) A few algorithms were tested, which included: Gaussian Naïve Baise; K-Means; Random Forests; Support Vector Machine. These were tuned to improve the results but Random Forests was selected as the one, which had the best performance.

4) Tuning the parameters of an algorithm can be used to improve its performance. Changing the parameters can be helpful to improve its predictive power or to train it better. If this is not done well the algorithm might not perform well with a new dataset. For this project, a RandomForestClassifier was used and the parameters tuned were the number of estimators and the “warm start” parameter was set to true.

5) For validation, a Stratified Shuffle Split was used. With this technique the algorithm was better trained and provided better results. A classic mistake during validation is to over fit the algorithm or to not train the algorithm well. With a Stratified Shuffle Split used to separate the data into several different train and test subsets, the algorithm could be re-

trained several times and in conjunction with the “warm start” parameter this allowed for better results.

6) The evaluation metrics used were Accuracy, Precision and Recall. For each of these metrics, the algorithm scored 96.1%, 91.4% and 71.6%, respectively. Accuracy is a metric that indicates that in 96.1% of the persons evaluated were correctly identified as a person of interest or not. Precision indicates that 91.4% of the persons who are of interest were identified correctly, while recall shows that 71.6% of those who are not a person of interest were confirmed to not be one.

Sources:

Tuning the parameters of your Random Forest model
https://www.analyticsvidhya.com/blog/2015/06/tuning-random-forest-model/

Feature Selection for Machine Learning in Python
http://machinelearningmastery.com/feature-selection-machine-learning-python/