MSc. Artificial Intelligence and Business Analytics

Python for Data Science

Final Project

Machine Learning to Identify Fraud in the Enron Corpus

1. Context and Enron Corpus

In late 2001, Enron, an American energy company, filed for bankruptcy after one of the largest financial scandals in corporate history. After the company's collapse, over 600,000 emails generated by 158 Enron employees - now known as the Enron Corpus - were acquired by the Federal Energy Regulatory Commission during its investigation. The data was then uploaded online, and since then, a number of people and organizations have graciously prepared, cleaned and organized the dataset that is available to the public today (a few years later, financial data of top Enron executives were released following their trial).

2. Project Description and Goal

Enron's financial scandal in 2001 led to the creation of a very valuable dataset for machine learning, on where algorithms were trained and tested to be able to find fraudulent employees, or persons-of-interest (POIs). In this project, a merged dataset of financial and email data will be used to go through the entire machine learning process.

The aim of this project is to apply machine learning techniques to build a predictive model that identifies Enron employees that may have committed fraud based on their financial and email data.

The dataset has:

- 14 financial features (salary, bonus, etc.),
- 6 email features (to and from messages, etc.)
- A Boolean label that denotes whether a person is a person-of-interest (POI) or not (established from credible news sources).

It is these features that will be explored, cleaned, and then put through various machine learning algorithms, before finally tuning them and checking its accuracy (precision and recall).

The objective is to get a precision and recall score of at least 0.42

First, the dataset will be manually explored to find **outliers** and **trends** and generally understand the data. Certain useful financial or email-based **features** will be chosen (manually and automatically using sklearn functions) and ensemble features created from those available, and then put through appropriate feature scaling. Then, numerous algorithms with **parameter tuning** will be trained and tested on the data.

The detailed results of the final algorithm, will be described in detail. The validation and evaluation metrics will be also shown and the reasoning behind their choice and its importance will be carefully explained. Finally, other ideas involving feature selection, feature scaling, other algorithms and usage of email texts will be discussed.

3. File Information

`/final_project/`:

- `poi_id.py`: Main file. Runs final feature selection, feature scaling, various classifiers (optional) and their results. Finally, dumps classifier, dataset and feature list **so anyone can check results**.
- `tester.py`: Functions for validation and evaluation of classifier, dumping and loading of pickle files.
- `my_classifier.pkl`: Pickle file for final classifier from `poi_id.py`.
- `my_dataset.pkl`: Pickle file for final dataset from `poi_id.py`.
- `my_feature_list.pkl`: Pickle file for final feature list from `poi_id.py`.

  Others function within `poi_id.py`

  - `feature_creation`: Functions for creating two new features
    - `'poi_email_ratio'` and `'exercised_stock_options'`.
  - `select_k_best` : Function for selecting k best features using sklearn's SelectKBest, sorts them in descending order of score.
  - `PlotOutliers`: Function for drawing plots of any two features colored by POI and non-POI.

`/tools/`:

- `feature_format.py`: Functions to convert data from dictionary format into numpy arrays and separate target label from features to make it suitable for machine learning processes.