# Enron Dataset

30/07/2020

This is final project report for my course "Introduction to Machine Learning" at Udacity.

The project helped me to understand and apply different machine learning algorithms I learned during my course.

## Background Information on EMail Data Set

Enron, which was deemed as powerful energy company, collapsed in the early 2000s. The reason Enron collapsed, was due to one of the greatest corporate fraud of all time.

## The project

The purpose of this project is to identify fraud from Enron Email datasets by comparing email exchanges between employees and company’s financial data. This project is to identify employees who might have been involved in the fraud and will be known POI or person of interest.

## Data Exploration

|  |  |  |
| --- | --- | --- |
| Measurement | Value | Observations |
| Total data points | 146 |  |
| POI data points | 18 |  |
| Low POI representation | around 12% |  |
| Non-POI data points | 128 |  |
| Number of features | 21 | 14 financial features, 7 email features |
| Missing values |  | All features have missing values |

## Handling Missing Values

Features with more than 50% missing values: deferral\_payments, loan\_advances ,restricted\_stock\_deferred, deferred\_income, long\_term\_incentive, director\_fees

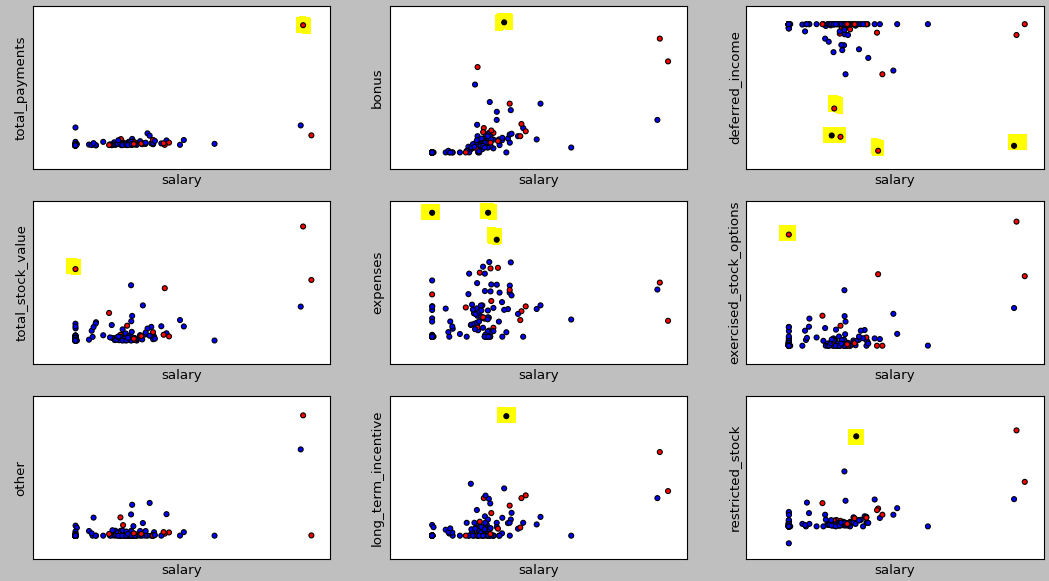
POIs with more than 50% missing values: deferral\_payments, loan\_advances, restricted\_stock\_deferred, director\_fees

Conclusion: The features restricted\_stock\_deferred, director\_fees, loan\_advances have both a large number of missing non-poi(>110) and poi(>17) values; therefore it would be imprudent to use them as features.

## Outliner Investigation

* Visualization was used for outlier detection. While removing the top 10% values would have been faster and 'programmatic', this was not appropriate here as it would remove a significant POIs from the dataset.
* A data point significantly stood out in the initial dataset, dwarfing the rest of the data. On inspection, this turned out to be the TOTAL which had been erroneously appended to the dataset.
* The resultant dataset was visualized by plotting salary against the other financial features (Figure 1). The email features were derived from the actual email dataset and therefore outlier detection was applied.

### Financial Outliners - Initial dataset



|  |  |  |
| --- | --- | --- |
| Feature | Outliers | Decision |
| Salary-Total Payments | 3 outliers in salary and total payments | The points correspond to key Enron employees (e.g.Chairman Kenneth Lay) thus we can safely assume that they are valid. |
| Bonus | A mid-wage non-POI LAVORATO JOHN J earns a higher bonus than the CEO! | Removed. |
| Deferred income | Five points below the average trend, of which 2 were non-poisALLEN PHILLIP K, FREVERT MARK A and 3 were pois HANNON KEVIN P, BELDEN TIMOTHY N, RICE KENNETH D | The points are retained. Removing 3 POIs would be detrimental for the quality of the data. |
| Total stock value, Exercised stock options | Though this appears to be a low-wage POI owning above-average stocks, on hindsight this, was caused due to a missing value in salary: HIRKO JOSEPH was the broadband chief of Enron. | While this data point is explainable with regards of stock, it was removed due to the large amount of missing values in it(including all the email data). |
| Expenses | Three mid-wage, non-POIs(SHANKMAN JEFFREY A, MCCLELLAN GEORGE, URQUHART JOHN A) incur huge expenses | "God-fearing souls splurging on rich stationery!" Removed. |
| Long term incentives | Non POI MARTIN AMANDA K | Removed |
| Restricted stock | Two mid-wage non-POIs(WHITE JR THOMAS E, PAI LOU L) | Removed |

## Financial Outliers after outliner removal

Following is much cleaner dataset.

