Enron Fraud Detection

(Enron Submission Free-Response Questions)

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? [relevant rubric items: “data exploration”, “outlier investigation”]

Answer:
Straight to the point, the goal of this project is to find the POIs out of enron's employees, using the the content that we have learned in udacity's classes, as machine learning techniques, problem solving and evaluation of the whole process and results. About the dataset and outliers detection, we have found something about 04 outliers: 'Total', 'The travel agency' , FREVERT MARK A', 'METTS MARK'.

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. (You do not necessarily have to use it in the final analysis, only engineer and test it.) In your feature selection step, if you used an algorithm like a decision tree, please also give the feature importances 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. [relevant rubric items: “create new features”, “intelligently select features”, “properly scale features”]

Answer:
Actually, we have used MinMaxScaler, tha transforms features by scaling each feature to a given range. This estimator scales and translates each feature individually such that it is in the given range on the training set, e.g. between zero and one. The transformation is given by:

In [ ]:

```
X_std = (X - X.min(axis=0)) / (X.max(axis=0) - X.min(axis=0))
X_scaled = X_std * (max - min) + min
where min, max = feature_range.
```

The transformation is calculated as:

In [ ]:

```
X_scaled = scale * X + min - X.min(axis=0) * scale
where scale = (max - min) / (X.max(axis=0) - X.min(axis=0))
```

This transformation is often used as an alternative to zero mean, unit variance scaling. As the POI's were taking larger amounts of money as bonus, in addition to their high salary so it can be stated that the ratio of bonus to salary of the POI's will be higher as compared to that of non-POI's. So i create a new feature called bonus-to-salary_ratio hoping that it may aid in the POI identification in the later parts of this project.

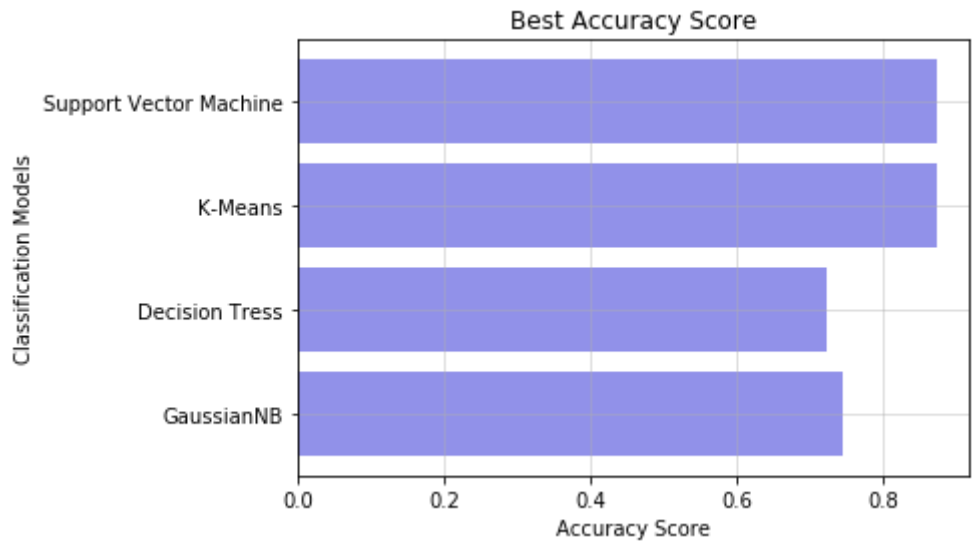
In [ ]:

```
df_clean['bonus-to-salary_ratio'] = df_clean['bonus']/df_clean['salary']
```

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”]

Answer:
Four algorithms have been choosed for this project:

1. Naive Bayes
2. Decision Tree
3. K-means
4. SVM



And, If we are talking about accuracy, the top 2 are K-Means and Support Vector Machine

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? (Some algorithms do not have parameters that you need to tune -- if this is the case for the one you picked, identify and briefly explain how you would have done it for the model that was not your final choice or a different model that does utilize parameter tuning, e.g. a decision tree classifier). [relevant rubric items: “discuss parameter tuning”, “tune the algorithm”]

Answer:
Tuning the parameters means adjusting the parameters of an algorithm so that it can handle a particular dataset better. Different parameter settings will result in different decision boundaries. If we don’t tune the parameters well, the algorithm won’t be able to generalize a dataset well and the final classification result might be less accurate.

5. What is validation, and what’s a classic mistake you can make if you do it wrong? How did you validate your analysis? [relevant rubric items: “discuss validation”, “validation strategy”]

Answer:
Validation techniques in machine learning are used to get the error rate of the ML model, which can be considered as close to the true error rate of the population. If the data volume is large enough to be representative of the population, you may not need the validation techniques. However, in real-world scenarios, we work with samples of data that may not be a true representative of the population. This is where validation techniques come into the picture.

5. 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. [relevant rubric item: "usage of evaluation metrics"]

Wikipedia says that, in pattern recognition, information retrieval and binary classification, precision (also called positive predictive value) is the fraction of relevant instances among the retrieved instances, while recall (also known as sensitivity) is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on an understanding and measure of relevance. In our case, as an example, the GaussianNB Precision is 12.50% and the GaussianNB Recall is 16.67%.