  Back to Data Analyst Nanodegree

Identify Fraud from Enron Email

* [**REVIEW**](https://review.udacity.com/)
* [**CODE REVIEW  2**](https://review.udacity.com/)
* [**HISTORY**](https://review.udacity.com/)

Requires Changes

**8 SPECIFICATIONS REQUIRE CHANGES**

Dear Student,  
Not a bad start to your project. However, some work is there to be done. There are a few changes to be made before this project can be approved as fully **Meets Specifications**.  
I have provided comments, recommendations, and suggestions for the sections that **Require changes**.

All the best for the next submission!

**Quality of Code**

**Code reflects the description in the answers to questions in the writeup. i.e. code performs the functions documented in the writeup and the writeup clearly specifies the final analysis strategy.**

Not a bad start. However, there are still a few updates to be made in the sections below.

**poi\_id.py can be run to export the dataset, list of features and algorithm, so that the final algorithm can be checked easily using tester.py.**

The **poi\_id.py** file runs with no errors and exports the dataset, list of features and algorithm successfully.

**Understanding the Dataset and Question**

**Student response addresses the most important characteristics of the dataset and uses these characteristics to inform their analysis. Important characteristics include:**

* **total number of data points**
* **allocation across classes (POI/non-POI)**
* **number of features used**
* **are there features with many missing values? etc.**

**Required**

The important characteristics are not being identified. Instead of uploading a link write the solutions and the figures.

MANDATORY

The report should contain the following information:

* The total number of data points
* The allocation between the POI and non-POI

(OPTIONAL)

* The total number of features
* There are missing values within the features. Which features contain missing values.

All of these characteristics can be found in the Project Overview section.

**To meet specifications**

* Provide the important characteristics of the data set as mentioned above. The report should at least mention the total number of data points and allocation between POI and Non-POI to meet specifications.

**Student response identifies outlier(s) in the financial data, and explains how they are removed or otherwise handled.**

**Optimize Feature Selection/Engineering**

**At least one new feature is implemented. Justification for that feature is provided in the written response. The effect of that feature on final algorithm performance is tested or its strength is compared to other features in feature selection. The student is not required to include their new feature in their final feature set.**

**Required**

The code script does show the new features being engineered and the report does give the justification for creating the new features. However, in order to meet the criterion, the report must conclude whether the newly created feature(s) are important or not. This can be done in two ways:

* Stating whether the new feature made it in the final list of important features
* Determining whether the newly created feature(s) had a positive or negative impact on Precision and Recall.

**Univariate or recursive feature selection is deployed, or features are selected by hand (different combinations of features are attempted, and the performance is documented for each one). Features that are selected are reported and the number of features selected is justified. For an algorithm that supports getting the feature importances (e.g. decision tree) or feature scores (e.g. SelectKBest), those are documented as well.**

**Required**

Good work here for using Decision Tree Classifier's feature\_importances\_ attribute to get the important features. The only thing the report is missing is in comparison to other methods considered to select the features why was Decision Tree Classifier preferred over methods such as Select K best ?

**If algorithm calls for scaled features, feature scaling is deployed.**

**Required**

The report needs to mention whether if scaling was used for this project or not. If scaling was used. The report must mention why and the method used. Finally, if scaling was not needed the report must also state why.

**Pick and Tune an Algorithm**

**At least two different algorithms are attempted and their performance is compared, with the best performing one used in the final analysis.**

**Response addresses what it means to perform parameter tuning and why it is important.**

**Required**

It is required to give a definition of Parameter Tuning in the context of Machine Learning and state the importance of parameter tuning.  
These articles should help with this:

* [Hyperparamter-tuning](https://en.wikipedia.org/wiki/Hyperparameter_optimization).
* [The-importance-of-paramter-tuning-](https://stackoverflow.com/questions/22903267/what-is-tuning-in-machine-learning).

**At least one important parameter tuned with at least 3 settings investigated systematically, or any of the following are true:**

* **GridSearchCV used for parameter tuning**
* **Several parameters tuned**
* **Parameter tuning incorporated into algorithm selection (i.e. parameters tuned for more than one algorithm, and best algorithm-tune combination selected for final analysis).**

**Validate and Evaluate**

**At least two appropriate metrics are used to evaluate algorithm performance (e.g. precision and recall), and the student articulates what those metrics measure in context of the project task.**

**Response addresses what validation is and why it is important.**

**Required**

Within the report explain what is Validation and why it is important in the context of Machine Learning.

RECOMMENDATION

This should help [What-is-Validation?](https://link.springer.com/referenceworkentry/10.1007%2F978-1-4419-9863-7_233)

**to meet specifications**

* Explain what is validation and why it is important

**Performance of the final algorithm selected is assessed by splitting the data into training and testing sets or through the use of cross validation, noting the specific type of validation performed.**

The verdict is correct here for using Stratified Shuffle Split as the validation method within the code script.

**Required**

However, the report does not mention that Stratified Shuffle Split was used as the validation method. To meet this criterion the report should mention:

* Why was Stratified Shuffle Split used? ( (HINT: Look at the allocation between POI and non-POI what does that tell us about the data!)
* How much percent of the data was split into training and data and testing data

**When tester.py is used to evaluate performance, precision and recall are both at least 0.3.**

Good work

DecisionTreeClassifier(class\_weight=None, criterion='gini', max\_depth=6,

max\_features=None, max\_leaf\_nodes=None,

min\_impurity\_decrease=0.0, min\_impurity\_split=None,

min\_samples\_leaf=1, min\_samples\_split=12,

min\_weight\_fraction\_leaf=0.0, presort=False, random\_state=42,

splitter='best')

Accuracy: 0.86207 Precision: 0.52166 Recall: 0.41550 F1: 0.46257 F2: 0.43313

Total predictions: 14000 True positives: 831 False positives: 762 False negatives: 1169 True negatives: 11238

 RESUBMIT PROJECT

[**DOWNLOAD PROJECT**](https://review-api.udacity.com/api/v1/submissions/1536494/archive)

[**2**  CODE REVIEW COMMENTS](https://review.udacity.com/)



**Best practices for your project resubmission**

Ben shares 5 helpful tips to get you through revising and resubmitting your project.

[Watch Video](https://review.udacity.com/)(3:01)