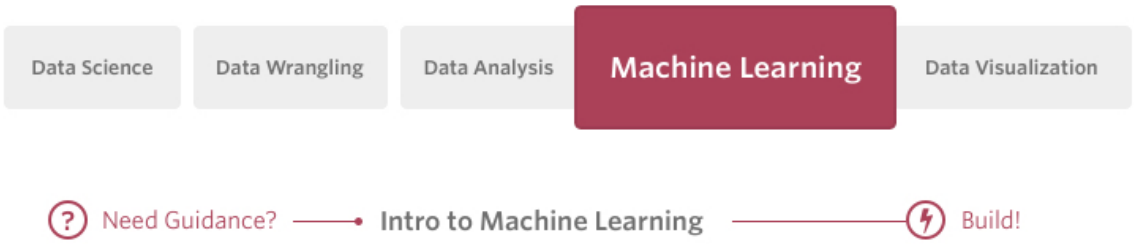
Identify Fraud From Enron Email



Project Overview

In this project, you will play detective, and put your machine learning skills to use by building an algorithm to identify Enron Employees who may have committed fraud based on the public Enron financial and email dataset.

Prepare for this project with: Intro to Machine Learning.

Note

If you have successfully completed the project for the Intro to Machine Learning course in the past (which entails having graduated from the course and having access to your course certificate), simply email us at dataanalyst-project@udacity.com with your passing evaluation and we'll give you credit for this project.

Why this Project?

This project will teach you the end-to-end process of investigating data through a machine learning lens.

It will teach you how to extract/identify useful features that best represents your data, a few of the most commonly used machine learning algorithms today, and how to evaluate the performance of your machine learning algorithms.

What will I learn?

By the end of the project, you will be able to:

- Deal with an imperfect, real-world dataset
- Validate a machine learning result using test data
- Evaluate a machine learning result using quantitative metrics
- Create, select and transform features compare the performance of machine learning algorithms
- Tune machine learning algorithms for maximum performance
- Communicate your machine learning algorithm results clearly

Why is this Important to my Career?

Machine Learning is a first-class ticket to the most exciting careers in data analysis today.

As data sources proliferate along with the computing power to process them, going straight to the data is one of the most straightforward ways to quickly gain insights and make predictions.

Machine learning brings together computer science and statistics to harness that predictive power.

This project is connected to the Intro to Machine Learning course, but depending on your background knowledge of machine learning, you may not need to take the whole thing to complete this project.

A note before you begin: the projects in the Intro to Machine Learning class were mostly designed to have lots of data points, give intuitive results, and otherwise behave nicely. This project is significantly tougher in that we're now using the real data, which can be messy and

doesn't have as many data points as we usually hope for when doing machine learning. Don't get discouraged--imperfect data is something you need to be used to as a data analyst! If you encounter something you haven't seen before, take a step back and think about a smart way around. You can do it!

Project Overview

In 2000, Enron was one of the largest companies in the United States. By 2002, it had collapsed into bankruptcy due to widespread corporate fraud. In the resulting Federal investigation, there was a significant amount of typically confidential information entered into public record, including tens of thousands of emails and detailed financial data for top executives. In this project, you will play detective, and put your new skills to use by building a person of interest identifier based on financial and email data made public as a result of the Enron scandal. To assist you in your detective work, we've combined this data with a hand-generated list of persons of interest in the fraud case, which means individuals who were indicted, reached a settlement, or plea deal with the government, or testified in exchange for prosecution immunity.

Resources Needed

You should have python and sklearn running on your computer, as well as the starter code (both python scripts and the Enron dataset) that you downloaded as part of the first mini-project in the Intro to Machine Learning course. You can get the starter code on git: *git clone*https://github.com/udacity/ud120-projects.git

The starter code can be found in the final_project directory of the codebase that you downloaded for use with the mini-projects. Some relevant files:

poi_id.py : starter code for the POI identifier, you will write your analysis here

final_project_dataset.pkl : the dataset for the project, more details below

tester.py : when you turn in your analysis for evaluation by your Udacity coach, you will submit the algorithm, dataset and list of features that you use (these are created automatically in poi_id.py). That coach will then use this code to test your result, to make sure we see performance that's similar to what you report. You don't need to do anything with this code, but we provide it for transparency and for your reference.

emails_by_address : this directory contains many text files, each of which contains all the messages to or from a particular email address. It is for your reference, if you want to create more advanced features based on the details of the emails dataset.

Steps to Success

We will provide you with starter code, that reads in the data, takes your features of choice, then puts them into a numpy array, which is the input form that most sklearn functions assume. Your job is to engineer the features, pick and tune an algorithm, test, and evaluate your identifier. Several of the mini-projects were designed with this final project in mind, so be on the lookout for ways to use the work you've already done.

The features in the data fall into three major types, namely financial features, email features and POI labels. financial features: ['salary', 'deferral_payments', 'total_payments', 'loan_advances', 'bonus', 'restricted_stock_deferred', 'deferred_income', 'total_stock_value', 'expenses', 'exercised_stock_options', 'other', 'long_term_incentive', 'restricted_stock', 'director_fees'] (all units are in US dollars) email features: ['to_messages', 'email_address', 'from_poi_to_this_person', 'from_messages', 'from_this_person_to_poi', 'poi',

'shared_receipt_with_poi'] (units are generally number of emails messages; notable exception is ‘email_address’, which is a text string) POI label: [‘poi’] (boolean, represented as integer)

You are encouraged to make, transform or rescale new features from the starter features. If you do this, you should store the new feature to my_dataset, and if you use the new feature in the final algorithm, you should also add the feature name to my_feature_list, so your coach can access it during testing. For a concrete example of a new feature that you could add to the dataset, refer to the lesson on Feature Selection.

Your submission will contain several files: the code/classifier you create and some written documentation of your work. We will evaluate your project according to the rubric here, only projects that satisfy all "meets expectations" items will pass. Please self-evaluate before you submit! If you don't think your project meets all the criteria, the project evaluator likely won't either.

Submission

Ready to submit your project? Go back to the portal, click on the project, and follow the instructions to submit!

- You can either send us a GitHub link of the files or upload a compressed directory (zip file).
- Inside the zip folder include a text file with a list of Web sites, books, forums, blog posts, GitHub repositories etc that you referred to or used in this submission (Add N/A if you did not use such resources).

It can take us up to 2 weeks to grade the project so keep checking back for updates.

If you are having any problems submitting your project or wish to check on the status of your submission, please email us at dataanalyst-project@udacity.com.

Items to include in submission:

Code/Classifier

When making your classifier, you will create three pickle files (my_dataset.pkl, my_classifier.pkl, my_feature_list.pkl). The project evaluator will test these using the tester.py script. You are encouraged to use this script before submitting to gauge if your performance is good enough. You should also include your modified poi_id.py file in case of any issues with running your code or to verify what is reported in your question responses (see next paragraph).

Documentation of Your Work

Document the work you've done by answering (in about a paragraph each) the questions found here.You can write your answers in a PDF, Word document, text file, or similar format. Include this document as part of your submission to the email address above.

Text File Listing Your References

A list of Web sites, books, forums, blog posts, github repositories etc. that you referred to or used in this submission (add N/A if you did not use such resources). Please carefully read the following statement and include it in your document “I hereby confirm that this submission is my work. I have cited above the origins of any parts of the submission that were taken from Websites, books, forums, blog posts, github repositories, etc.

Good Luck!