## **Datasets** (and their respective Credits/Acknowledgements)

*Name to identify which dataset*

*Year to evaluate dataset’s outdatedness (& accuracy)*

***Ideal domains to cover: phone message (& number and maybe embedded links), urls (& web features) , emails (& email domain), ~~credit card~~***

*Visit their respective websites to view contributors for acknowledgements*

| **Dataset Name:** | **creditcard1.csv** |
| --- | --- |
| Year of Collection: | 2013 |
| Description: | The dataset contains transactions made by credit cards in September 2013 by European cardholders.  This dataset presents transactions that occurred in two days, where we have 492 frauds out of 284,807 transactions. The dataset is highly unbalanced, the positive class (frauds) account for 0.172% of all transactions.  It contains only numerical input variables which are the result of a PCA transformation. Unfortunately, due to confidentiality issues, we cannot provide the original features and more background information about the data. Features V1, V2, … V28 are the principal components obtained with PCA, the only features which have not been transformed with PCA are 'Time' and 'Amount'. Feature 'Time' contains the seconds elapsed between each transaction and the first transaction in the dataset. The feature 'Amount' is the transaction Amount, this feature can be used for example-dependant cost-sensitive learning. Feature 'Class' is the response variable and it takes value 1 in case of fraud and 0 otherwise. |
| Website: | <https://www.kaggle.com/mlg-ulb/creditcardfraud> |

| **Dataset Name:** | **website.csv (PLANNED TO USE)** |
| --- | --- |
| Year of Collection: | 2015 - 2017 |
| Description: | This dataset contains 48 features extracted from 5000 phishing webpages and 5000 legitimate webpages, which were downloaded from January to May 2015 and from May to June 2017. An improved feature extraction technique is employed by leveraging the browser automation framework (i.e., Selenium WebDriver), which is more precise and robust compared to the parsing approach based on regular expressions. |
| Website: | <https://www.kaggle.com/shashwatwork/phishing-dataset-for-machine-learning> |

| **Dataset Name:** | **Email.csv (PLANNED TO USE)** |
| --- | --- |
| Year of Collection: | 1998 - 2007 |
| Description: | I have used some part of the (<https://www.kaggle.com/rtatman/fraudulent-email-corpus/home>) dataset to create this new data set for supervised learning.  This dataset is a collection of more than 2,500 "Nigerian" Fraud Letters, dating from 1998 to 2007.  These emails are in a single text file. Each e-mail has a header which includes the following information:   * Return-Path: address the email was sent from * X-Sieve: the X-Sieve host (always cmu-sieve 2.0) * Message-Id: a unique identifier for each message * From: the message sender (sometimes blank) * Reply-To: the email address to which replies will be sent * To: the email address to which the e-mail was originally set (some are truncated for anonymity) * Date: Date e-mail was sent * Subject: Subject line of e-mail * X-Mailer: The platform the e-mail was sent from * MIME-Version: The Multipurpose Internet Mail Extension version * Content-Type: type of content & character encoding * Content-Transfer-Encoding: encoding in bits * X-MIME-Autoconverted: the type of autoconversion done * Status: r (read) and o (opened) |
| Website: | <https://www.kaggle.com/llabhishekll/fraud-email-dataset> (data) OR  <https://www.kaggle.com/rtatman/fraudulent-email-corpus/home> (original, un-EDA-ed) |

| Dataset Name: | **enron**v1.csv (not sure if full) **(PLANNED TO USE)** |
| --- | --- |
| Year of Collection: | 2015 version |
| Description: | The Enron email dataset contains approximately 500,000 emails generated by employees of the Enron Corporation. It was obtained by the Federal Energy Regulatory Commission during its investigation of Enron's collapse.  This is the May 7, 2015 Version of dataset, as published at <https://www.cs.cmu.edu/~./enron/> |
| Website: | <https://www.kaggle.com/wcukierski/enron-email-dataset> |

| Dataset Name: | **enron**v2 (use pd import coz files are kinda large, or you can download on your own via website below) **(PLANNED TO USE)** |
| --- | --- |
| Year of Collection: | 2015 version |
| Description: | import pandas as pd  df = pd.read\_csv('<https://query.data.world/s/auzeqhtwaa27u45dcupbakhczxvpw6>')  df = pd.read\_csv('https://query.data.world/s/5jrhpcgqrvq7oiutvhidyluhvo2zio')  df = pd.read\_csv('https://query.data.world/s/pggnkxiptarzp3uixtl7mtuqfctsac')  df = pd.read\_csv('https://query.data.world/s/6ykzyk7l4uz7pqdy7lex7z2n7fb3fk')  df = pd.read\_csv('https://query.data.world/s/p3wqohjs5kzcgvprg67cjhkdjrgjbt')  df = pd.read\_csv('https://query.data.world/s/ogc5mioavpcx3imokrplogtpnrkyir')  df = pd.read\_csv('https://query.data.world/s/rol4dzbh7wr26fz3nasnppjptbudkd') |
| Website: | <https://data.world/brianray/enron-email-dataset> |

| Dataset Name: |  |
| --- | --- |
| Year of Collection: | Can get the most recent email 2021 |
| Description: |  |
| Website: | https://monkey.org/~jose/phishing/ |

| Dataset Name: |  |
| --- | --- |
| Year of Collection: |  |
| Description: |  |
| Website: |  |

| Dataset Name: |  |
| --- | --- |
| Year of Collection: |  |
| Description: |  |
| Website: |  |