CSE 519 -- Data Science (Fall 2021)
Prof. Steven Skiena
Homework 2: Exploratory Data Analysis in iPython
Due: Thursday, September 23, 2021 (9:45 AM)

This homework will investigate doing exploratory data analysis in iPython. The goal is to get you fluent in working with the standard tools and techniques of exploratory data analysis, by working with a data set where you have some basic sense of familiarity.

This homework is based on Microsoft Malware Prediction at Kaggle, revolving around predicting the probability of a device being infected by malwares. More than just data exploration, you must also join the challenge and submit your model, to get a score from Kaggle. You are to explore the data and uncover interesting observations about antivirus products and their effectiveness. You will need to submit your code files in three different formats (.ipynb, .pdf and .py). Make sure to have your code documented with proper comments and the exact sequence of operations you needed to produce the resulting tables and figures. The submission steps are discussed below.

Data downloading

First of all, you need to join the challenge and download the data here. The description of the data can also be found at this page.

Python Installation

Instead of installing python and other tools manually, we suggest installing **Anaconda**, which is a Python distribution with a package and environment manager. It simplifies a lot of common problems when installing tools for data science. More introduction can be found here. Installation instructions can be found here.

Another option can be using **Google Colaboratory.** This is another option for those who want to run their Jupyter notebook remotely instead of installing the required packages locally. Colab allows you to write and execute Python in your browser, with

- Zero configuration required
- Free access to GPUs
- Easy sharing

If you are an expert of Python and data science, what you need to do is install some packages relevant to data science. Packages that I believe you will definitely use for this homework include:

- pandas
- scikit-learn
- numpy

- Matplotlib
- seaborn

The google colab notebook contains boilerplate code to download the data to your google drive and a dictionary containing the features along with its data type. Make a copy of the notebook before you start your HW. Besides this, we also provide a list of 40 features in a list named *use_cols* which you will use for all the questions. *Only for questions 5 and 8 you are free to explore features that are not in this list.*

Tasks (100 pts)

1. Load the CSV into a dataframe with the features as listed in the *use_cols* variable. (5 points)
2. There are two parts for this task:
    a. Define a measure of computer power as a function of RAM, processor core count and any other relevant features you find. What is the distribution of power among the machines in the dataset? (10 points)
    b. Are powerful computers more or less likely to have malware than underpowered machines? Plot power vs malware detection to support your conclusion. (5 points)
3. Software is updated to fix vulnerabilities when found. But these updates can also open a can of worms. Produce plots showing the number (and %) of malware detections against Census_OSBuildNumber and also against Census_OSBuildRevision. Discuss what you find. (10 points)
4. Investigate the question of whether antivirus software(s) reduces the amount of malware. Does the number of antivirus products you use matter? What is your conclusion and what is your evidence supporting it? (10 points)
5. Create **3** plots of your own using the dataset that you think reveal something very interesting. Explain what it is, and anything else you learned from your exploration. (15 points)
6. Now build a baseline model for this task. We will call this *Model 0.* You will train a logistic regression model on 80% of the training data and test it on the remaining 20% *chosen at random*. List the features used and print the error rate along with the AUC score of this model. What do you make of the error rate? (10 points)
   [Note: This baseline model does not need to be very good/sophisticated. They are usually used for checking out some preliminary ideas and their performance.]
7. There are two parts for this task:
    a. *Cleaning Features*: Features can be preprocessed to improve them before feeding into the model (e.g. normalize or scale the input vector, convert non-numerical value into float, or do a special treatment of missing values, etc). This can significantly improve the performance of your model. Do preprocessing for the features. Explain what you did. (15 points)

b. *Final Model Creation:* Create two models.
    i. *Model 1* should use the cleaned features (All of the features do not have to be preprocessed) and logistic regression for training.
    ii. *Model 2* should use the cleaned features (All of the features do not have to be preprocessed) and an algorithm other than logistic regression (e.g. Random Forest, Nearest Neighbor, etc) for training.
    [Note: *scikit-learn* is a user-friendly library which is used to perform data loading, pre-processing, transformations, algorithms and metrics needed for Data Science and Machine learning]

    Use the same training and test set you used in question 6. Report the error rate for Model 0, 1, and 2 in a table. Compare their performance and explain your reasoning for the differences in their performances. (10 points)

8. Write the probability of detection for the test instances (*test.csv*) into a csv file as shown in 'sample_submission.csv' at Kaggle. Submit this for every model you develop to the competition website. Report the private and public score for your best submission along with the number of submissions. Include a snapshot of your best score on the website as confirmation. Be sure to provide a link to your Kaggle profile. (10 points)

Be honest. This is your first modelling experience, and I am hoping to see you learned something, not just where you are ranked on the leaderboard.

Rules of the Game

This assignment must be done **individually by each student**. It is not a group activity.

1. If you do not have much experience with Python and the associated tools, this homework will be a substantial amount of work. Get started on it as early as possible!
2. All of your written responses should be put in the appropriate place in your notebook template. Get the template notebook form from Google Classroom!! Or from here
   You are allowed to add more cells, but definitely fill out the cells we give.
3. We will discuss topics like logistic regression in detail only after the HW is due. Muddle along for now, and we will understand the issues better when we discuss them in the course.
4. To ensure that you are who you are when submitting your models, have your Kaggle profile show your face as well as a Stony Brook affiliation.
5. There are some public discussions and demos relevant to this problem on Kaggle. It is okay for students to read these discussions, but they must write the code and analyze the data by themselves.
6. You will submit your code so we can run it through MOSS to detect copying and plagiarism. Do your own work!!

7. Our class Piazza account is an excellent place to discuss the assignment. Check it out at piazza.com/stonybrook/fall2021/cse519.

Submission

Submit everything through Google classroom. As mentioned above, you will need to upload:

1. The Jupyter notebook all your work is in (.ipynb file), derived from the provided template
2. Python file (export the notebook as .py)
3. PDF (export the notebook as a pdf file)

These files should be named with the following format, where the italicized parts should be replaced with the corresponding values:

1. cse519_hw2_*lastname_firstname_sbuid*.ipynb
2. cse519_hw2_*lastname_firstname_sbuid*.py
3. Cse519_hw2_*lastname_firstname_sbuid*.pdf