CSE 519 -- Data Science (Fall 2019)
Prof. Steven Skiena
Homework 2: Exploratory Data Analysis in iPython
Due: Thursday, September 26, 2019 (9:59 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 IEEE Fraud Detection on Kaggle, revolving around predicting the probability of whether a transaction is fraudulent. More than just data exploration, you must also join the challenge and submit your model before the deadline, to get a score feedbacked from Kaggle. You are to explore the data and uncover interesting observations about customer transactions. You will need to submit all your results in a single Google form and your code files in three different format (.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 have been 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 to install **Anaconda**, which is a Python distribution with 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.

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

Another modern alternative is Google Colaboratory. This is another option for those who want to run their Jupyter notebook remotely instead of installing the required packages locally.

Tasks (100 pts)

This data set is large with many fields. Parts 1 to 5 will be restricted to the following non-anonymous fields for simplicity:

- TransactionID
- DeviceType (mobile/desktop/...)
- DeviceInfo (Windows/MacOS/…)
- TransactionDT (time delta from reference)
- TransactionAmt (amount in USD)
- ProductCD (product code - W/C/H/R/...)
- card4 (card issuer)
- card6 (debit/credit)
- P_emaildomain (purchaser email)
- R_emaildomain (recipient email)
- addr1 / addr2 (billing region / billing country)
- dist1 / dist2 (some form of distance - address, zip code, IP, phone, …)

The target we are trying to predict is:

- isFraud (provided as boolean whether transaction was fraud or not, predict as probability of fraud)

1. Filter out your data to examine just the fraudulent transactions. For each field above, examine the distribution of the values, and explain any interesting insight you get from this. How do do the distributions on fraudulent transactions compare to the non-fraudulent ones? (15 points)
2. The **addr2** field gives a code (but not name) associated with the country of the purchaser. **TransactionDT** shows the time passed from some reference for each transaction. By looking at the time of day of the transactions, we can infer what waking hours are associated with the country relative to the reference time. Analyze the frequency distribution of transactions by time for the most frequent country code, as per the **addr2** field. Plot this distribution. Explain your findings. (15 points)
3. **ProductCD** refers to a product code. Make your best educated guess on which codes correspond to the most expensive products and which to the cheapest products. Justify with analysis. (10 points)
4. Plot the distribution between the time of day and the purchase amount. What is the correlation coefficient? Note that some cleaning is necessary to get a meaningful time of day. (10 points)
5. Create a plot of your own using the dataset that you think reveals something very interesting. Explain what it is, and anything else you learned. (15 points)

6. Now, try to build a prediction model that works to solve the task. You are allowed to use additional variables from the dataset if you wish. Perhaps it will use linear regression. Perhaps it will preprocess features (e.g. normalize or scale the input vector, convert non-numerical value into float, or do a special treatment of missing values). Perhaps it will use a different machine learning approach (e.g. nearest neighbors, random forests).

   Explain what you did. What is the accuracy of your model? (20 points)

7. Predict all the fraud cases for the test instances at file “sample_submission.csv”. Write the result into a csv file and submit it to the website. You should do this for every model you develop. Report the rank, score, number of entries, for your highest rank. Include a snapshot of your best score on the leaderboard as confirmation. Be sure to provide a link to your Kaggle profile. (15 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 will be put in the appropriate place in your notebook template. Get the template notebook form from Google Classroom!! You are allowed to add more cells, but definitely fill out the cells we give.
3. We will discuss topics like linear 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/fall2019/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