CSE 519 -- Data Science (Fall 2018)
Prof. Steven Skiena
Homework 3: Data Integration and Modeling
Due: Tuesday, October 23, 2018

This homework will investigate data integration and model building in IPython. It is based on the Google Analytics Customer Revenue Prediction Kaggle challenge, revolving around predicting how much the GStore customer will spend. You are charged with predicting the natural log of the sum of all transactions per user. For every user in the test set, the target will be:

$$y_{user} = \sum_{i=1}^{n} transaction_{user_i}$$

$$target_{user} = \ln(y_{user} + 1)$$

The dataset has the following fields:

fullVisitorId- A unique identifier for each user of the Google Merchandise Store.
channelGrouping - The channel via which the user came to the Store.
date - The date on which the user visited the Store.
device - The specifications for the device used to access the Store. This field is a JSON with additional information such as browser, operatingSystem, deviceCategory etc.
geoNetwork - This section contains information about the geography of the user. This JSON field contains subcolumns such as continent, country, region etc.
sessionId - A unique identifier for this visit to the store.
totals - This section contains aggregate values across the session. This is also a JSON field with a column transactionRevenue whose value is to be predicted.
trafficSource - This section contains information about the Traffic Source from which the session originated.
visitId - An identifier for this session. This is only unique to the user. For a completely unique ID, you should use a combination of fullVisitorId and visitId.
visitNumber - The session number for this user. If this is the first session, then this is set to 1.
visitStartTime - The timestamp when the session started

Many of the tasks mirror those of the previous assignment, as practice makes perfect. As in the previous assignment, 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.

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.

Tasks (100 pts)

1. Take a look at the training data. There may be anomalies in the data that you may need to factor in before you start on the other tasks. Make a note of the anomalies that you notice. Clean the data first to handle these issues. Explain what you did to clean the data (in bulleted form). (10 points)
2. Generate a heatmap and two other plots (with a subset of variables) visualizing interesting positive and negative correlations. Explain the reason for your choice for these variables and any interesting results associated with them. (15 points)
3. Cluster the data based on geographic information available with a subset of variables that you find relevant. Include a visualization plot. Describe your inferences from the clustering and discuss their significance. (15 points)
4. Define a buying score or probability function for each user, which predicts the likelihood of a user buying a product from the GStore. Rank the ten most likely users as who will buy a product from the store. Does it seem that you that it produces good results? Report why or why not. (15 points)
5. Identify at least one external data set which you can integrate into your transaction prediction analysis to make it better. Discuss/analyze the extent to which this data helps with the prediction task. (10 points).
6. Finally, build the best prediction model you can to solve the Kaggle task. Use any data, ideas, and approach that you like. Submit the results of your best models on Kaggle. Report the rank, score, number of entries, for your highest rank. Include a snapshot of your best score on the leaderboard as confirmation. (20 points)
7. Do a permutation test to determine whether your model really benefits from each input variable you use. In particular, one at a time, for each relevant input variable, permute the value of this variable and see how they impact the accuracy of the results. Run enough permutations per variable to establish a *p*-value of how good your predictions of log of sum of transactions per user are. You can use whatever metric you wish to score your model (like mean absolute error). (15 points)