Data Science Exercise

Background

At ***, our mission is to improve every human interaction via data science. We have a lot of data. Our customers send us billions of events representing user and customer behaviors (e.g. logging into an app, buying something, visiting a website, etc.) a month via our API that we validate, normalize and store in realtime. We store and aggregate data in a datastore that allows us to execute low latency queries with high throughput -- so asking questions is fast. On top of that datastore, we're now starting to build algorithms to test hypotheses and discover patterns in the data.

For this exercise, we're going to predict future purchasing from a history of customer transactions.

The Exercise

Attached to this doc is a CSV file representing approximately 50,000 transactions from an ecommerce store that sells widgets. The data has the following format:

CustomerID	**Timestamp**	**PurchaseValue**
hash_value	2017-01-01T00:00:00Z	100.00

Each row represents a purchase. The first column is the **CustomerID** and it's possible to have multiple rows with the same **CustomerID**. The second column is a UTC timestamp of the transaction and the third column is the monetary value of the transaction.

To do this exercise, use the modified beta geometric and the gamma gamma model in the lifetimes python library (https://github.com/CamDavidsonPilon/lifetimes). Consider the work of Fader and Hardie (cited in the python package) while implementing your model. Keep in mind there are very advanced models that could take weeks and months to tune, but for this exercise, stick to what has already been implemented in the lifetimes library.

We want to see your best work - your submission should be well organised, readable, and should show an understanding of the model of customer purchasing behavior.

Questions

Please answer the following **8 questions** and attach the **jupyter notebook** (.ipynb) you use to perform your analysis. Put your name in your notebook's file name.

Warmup/basic python:

1. Implement the modified BG model from the lifetimes package using the data we provide.

2. List the 100 customers predicted to make the most purchases over the next 12 months.

3. List the 100 customers predicted to spend the most over the next 12 months.

Analysis:

4. Explain the statistical assumptions that the model makes. Do you think those assumptions are valid for a model of a customer buying widgets from an ecommerce store?

Write a simulation:

5. Write a simulation that shows how many customers are alive after 10 days, 1 year, 10 years and 100 years and how many purchases they have made in that time using the modified BG model. Use a **simple random sample** of 100 customers and show the results for 1 run of your simulation. We've outlined a possible approach below:
    a. Generate a random sample of 100 customers.
    b. Simulate how each customer makes purchases over time.
    c. Count how many purchases the customers have made in 10 days.
    d. Count how many customers are alive after 10 days.
    e. Repeat b-d for 1 year, 10 years, 100 years.
    f. Tip: Your simulation should show a different number of customers alive each time the code is evaluated. (Often it is interesting to look at this range of values, but for this exercise, we are only looking for you to show us the result of one run.)
    g. Tip: You will have to write code beyond what is available in the python package.

Analysis of results of questions 1-5:

6. The model works well in some ways and poorly in others for the data we provide. Explain what it does a well and what it does poorly.

7. What are one or two ideas that could address the deficiencies of the model? (No need to implement, just describe.)

Explain, independent of questions 1-7:

8. Suppose the intended use of the modified BG model is binning customers into one of the following categories: highly likely to purchase in the next year, somewhat likely to purchase in the next year, unlikely to purchase in the next year, highly unlikely to purchase in the next year. Suppose you want to compare this model with a model your colleague is proposing. How would you evaluate the two models? Do not implement, but do show a mockup of the results of your evaluation.