Problem formulation

Assume we are a mail-delivery book startup. You fill out a profile with us about your book preferences, and we will send you 5 books each month for free. Whatever you like you keep and pay for, and whatever you don't like you send back.

As a company we have to buy our books ahead of time. We took out a loan last month to buy our original batch of books. The value of the loan was the total cost of all the books that we bought. We made some money back through customers buying our books last month. Next month, we know which books we will be sending to which customers, but we do not know who will buy what books. It costs 0.60/book each way for shipping books to and from customers. The question for you to answer is:

Will we be able to both pay back our loan and afford our next book purchase order?

Data

We've provided some data for you to help answer this question. All of the following are csv files.

- original_purchase_order: Quantity of books originally purchased indexed by a unique product ID.
- next_purchase_order: Planned quantity of books to purchase at the end of next month indexed by a unique product ID.
- customer_features: Customer features generated by the profiles that customers fill out with us indexed by a unique customer ID.
- product_features: Product features that we have generated from our knowledge of our books indexed by a unique product ID.
- last_month_assortment: Data on which books were sent to which customers and whether or not the customer purchased the book. There is no index.
- next_month_assortment: Data on which books will be sent to which customers next month. There is no index.

Purpose

The purpose of this test is to assess how well you translate business problems into machine learning solutions as well as get an understanding of your coding ability and style. This is also a chance to give you insights into our business and some problems that we may be thinking about.

This test serves as a minimum bar that we set for interviewing candidates. You do not get extra consideration for going above and beyond the requirements of the test because we do not wish to favor candidates with infinite free time.

We do not want to assess your feature engineering prowess or how accurate you can make your machine learning model because we do not want you to spend all day on this.

With that being said, you **should build valid features** and **have solid reasoning** behind the parameters that you have chosen for your model. We are trying to be respectful of your time. We are not assessing how elaborate your script can be - we do not expect logging or anything like that.

We expect you to spend no more than 3 hours on this assessment.

Expectations

Your responses to this assessment should be a script that can be run on the command line, reads in the necessary data files as arguments to the script, runs the necessary calculations, and finally prints 'Yes' or 'No' to the question of whether or not we can pay back our loan and afford the next book purchase order.

Below is an example of how this might be run:

```
$ python are_we_going_to_survive.py original_purchase_order.csv
next_purchase_order.csv customer_features.csv product_features.csv
last_month_assortment.csv next_month_assortment.csv
```

Please write your script in Python. We use Python for all of our internal data science systems and would like you to be able to integrate yourself as quickly as possible should you start working here. Try to make your code as clean as possible .You should create some sort of machine learning model for answering the question (as opposed to simply looking at average conversion rate or something like that). However, we do not expect you to build models from scratch. numpy, scipy, scikit-learn, and everything else is all fair game.

Lastly, please write a brief description explaining:

- Which machine learning model you chose to use and why
- How you validated your model and why you chose to validate it in this way.
- How to run your script
- Any other features you would want to use (other data points you want to capture)
- Brief variables you would want to engineer