Numerator Data Science Assessment

1. Describe a possible approach you would take to solve for missing receipts in the system?

The goal for using this approach is the explain the gap between the benchmark and the receipts that are being submitted.

1. Estimate the projected contribution of the individual to the benchmark
2. Estimate the effects of the point of sales variables to the final
    a. There are many ways to estimate effects but the quickest method to provide proof of concept would be a boosting model.
3. Apply these estimates of effects to the existing data on the collected receipts and use that to forecast projections.

What is the projected total sales for the year.

Taking the sum of a 12 year forecasted monthly predictions. The estimated projected sales for the time period between 585,310,113.

This number came from taking the sales predictions for 9/21 - 10/22 and summing up the monthly projected sales values.

What is the projected household penetration for the year?

In order to estimate the projected household penetration, the following metrics were calculated.

1. The number of submissions as a proportion of benchmark sales. Using the number of submissions, the items purchased, and the total spent as a portion of the benchmark.
2. Calculated the sales proportion divided by the coverage rate of the sales monthly.
3. Divide the above metric by the metric.

The estimated projected household penetration for the year is .3% of US households.

What percentage of panelists receipts do we actually collect?

The coverage rate is a good proxy for the percentage of panelists receipts that is actually collected.

The coverage takes into account the number of users as well as the sales recorded in comparison to the benchmark.

Therefore the average receipt coverage rate is about 2.5%

How might we scale the process and what are the pitfalls.

1. The process can be scaled by creating a forecast that is run monthly and produces new estimates and predictions.
    a. This is advantages as it takes into account the completed month and adjusts projections with continued updated data.
    b. This allows a more accurate upper and lower bound for prediction intervals.
    c. With the frequency of re-running the model, the model will only learn faster and better
2. Pitfalls
    a. Putting a boosting model as well as a time series forecast into production is a very computationally heavy job and make either take a very long time to fun, or require external toolage (EC2, Sagemaker).
    b. Manual adjustments may be needed during unprecedented changes to the model which make cause a time discrepancy in producing the results.