Questions

Also include a word document with your submission addressing each of the following:

- Why did you choose your prediction methodology and how did you assess performance?

  I realized the problem is not a simple regression problem and we are dealing with a time series problem because visit volume and screening count are both correlated to each other and are dependent on time.

  Among the time series algorithms, I picked the Vector AutoRegression algorithm that can handle bi-directional data. Univariate time series algorithms are not suitable for this problem since screening and visit volume are correlated.

  For validation, I split the data into train and test and kept the last 2 weeks of the data for testing. The predictions of the model versus the actuals looked very promising. I also looked at the RMSE and correlation between the prediction and actual and they looked good.

- What are the strengths and weaknesses of your approach? Be sure to include important model assumptions.
  1. The VAR model assumes the input data is stationary, however I suspect the data that we are dealing with is seasonal. Also, I did not test for stationarity in the code, and did not include codes to make the data stationary.
  2. Model also assumes causation amongst the time series. The basis behind Vector AutoRegression is that each of the time series in the system influences each other. That is, we can predict the series with past values of itself along with other series in the system.
  3. We did not have enough data to train a good model, we only had access to 1 season of screening count so it is expected that the model won't predict well for the unseen seasons.
- How confident are you in the daily prediction results?

  I am fairly confident in the daily predictions with the amount of information we provided to the model. I did not take out the first 3 months of data where the screening count is 0, so the model can pick up the trend for number of visits, I think the model would over predicts the screening counts by a little bit but for the purpose of this problem that might be ok.

  However maybe there is a better approach to get better predictions by using past years information on the visit counts and maybe building separate models for number of visits and using that in the screening capacity forecast.

- Provide 2 questions you would ask the stakeholders
  1. Do you prefer to overpredict or underpredict the screening capacity?
  2. Do you have resources to retrain the model daily to improve the prediction performance?
- Given the output of your prediction, what recommendation would you make to the stakeholders?
  - The current model is overpredicting mainly because we included few months of 0 screening count data

- We might be able to find a better approach by doing some research, so if you have resources for couple of hours research that might help us to improve the prediction performance significantly
- If you retrain the model on a daily basis you can increase the accuracy by providing more information to the model.