This report is to summarize some of my findings. The details or intermediate steps are in the jupyter notebook.

1. First I will outline some work done on the data cleaning side.
    a. The 'Signal' data has a series of 0 value data towards the end. And the 'Adj Close' data for the ETF has some negative data. The 'Adj Close' prices was picked as they are adjusted for stock splits and dividend payments. Therefore, it reflected the 'true' stock price and can be used to compute the returns. I have performed filtering with the z-score. Any data with z-score greater than 5 would be considered outliers as it is very rare with about 1000 data points. I performed two step filtering and correction:
        i. First with the raw data itself: 8 outliers are identified. 6 of these records are 0s for signal in most recent days. I cannot figure out good ways to correct them. If I just fill-forward, I may end up with a series of 0 returns, which may not be realistic. So removal may be good to avoid any contamination of the data. For the other two outliers: 1) the signal one – I corrected with the previous day's value; 2) the 'adj close' big negative one – it should be positive. If I change the sign, I end up with same return as the 'close' prices.
        ii. Then compute the returns and correct those with |z-score|>=5. It is a bit tricky in this case as one abnormal data point can result two abnormal returns: one for today, the other for the next day. So I wrote a loop to handle each record sequentially. This way the corrected ones will not result abnormal returns in the next day.
            1. The correction for signal and 'adj close' price is performed differently.
                a. For signal, I just corrected/replaced it with the previous day's price which avoids look-ahead bias as I did before. There can be more complicated ways. Since there are not many records for correction, this simple way might work as well.
                b. For 'adj close', I found that the corresponding 'close' returns are normal. Since in most cases the two returns are in sync, I just use 'close' price return to back out today's price.
            2. Here total 6 records got corrected: 4 for signal and the other 2 for 'adj close'.
    b. There is a block of duplicated data in the 'adj close' prices, 12 in total. They are obviously not correct. I used the 'Open' 'High' 'Low' prices to help correct them as their prices keep on changing every day: I took the average return of the three prices. And from this estimated return I backed out the price data.
    c. There is another 'adj close' data with 3 duplicates. Since two of them are consecutive, so I correct the second one with the similar algorithm as before.
    d. For signal, there are also 3 or 2 duplicates exists. Since I do not have good way to backout, I just leave them there as they are mostly not consecutive.

2. I performed two stages in forecasting the ETF price to understand its efficacy.
    a. First, I used the raw price itself as the target and the signal as the explanatory variable to run linear regression. For the data, I split it 80:20 for training (in sample or IS) and test (out of sample or OOS) purposes. The signal has good R2 in sample but shows no predictive power OOS (with large MSE and negative R2).
        i. To rule out the possibility that the data for training may not be relevant. I also tested with various trailing periods from 30 to 210 days to predict OOS 5 day price. The conclusion is similar judging from OOS R2.
        ii. Since the ACF and PACF plots show that the ETF prices follow strong AR(1) process. I used the 'Adj Close' lag 1 day as the predictor. It has much better IS and OOS R2. This comparison shows that the signal doesn't have much predictive power in predicting next day's stock price.
    b. Next, I move forward to experiment with the daily returns as the returns are of much interest for portfolio managers or traders.
        i. Both the signal (returns) and the lag 1 (returns) models do not have good results in the training period. The p-values are not statistically significant.
        ii. Then I think that the signal might have some predictive power under some special market environment. For this study, I performed the following:
            1. Perform cluster analysis on the signal data to have groups of data points with different characteristics.
            2. Run linear regression for data points in each of those groups and output their IS R2.
            3. Identify a particular cluster/group with significant R2 and dig further: 1) check the p-value with the regression is statistically significant: the answer is YES; 2) run OOS stats, i.e., root mean square error (RMSE) and R2. Unfortunately, the OOS measures do not look good.
            4. Then I switched to a different type of regressor, i.e., decision tree based regressor. The results are promising: R2 are positive both IS and OOS. However, since there are so few data points in the OOS period that satisfies this cluster characteristics, it may not be convincing.
            5. I relaxed the condition to identify this cluster. Basically, this cluster is with big negative returns for the signal. I instead used the z-score (threshold set at -1.6) to pick up this regime to have more (actually doubled) data points. The IS and OOS results look good this time. The IS and OOS R2 are 0.42 and 0.41 respectively. The RMSE in IS and OOS periods are: 0.9% and 1.0%. Both of these measures look consistent.
            6. To verify the signal has some predictive power in this special regime. I also run regression with the lag 1 return. It is not statistically significant even in the IS period.

3. Summary for the portfolio manager (p.m.).
   From the above study, I can conclude that the signal does not have much predictive power under normal market conditions. However, it might work under some market environment where the signal has big negative returns.
   a. To verify this, we need to understand how the signal is constructed and whether it makes sense to have such efficacy if at all.
   b. Data quality is very important. We need to check with the vendor what is the actual prices of the signals on those days we identified as outliers.
   c. We need more data to verify the efficacy. I think the data points in the 'big down' market environment for test are not sufficient to draw a conclusion. Since the first quarter of 2020 has experienced large market downturn, it can be a very good period to put the signal into (stress) test.

4. Below are some questions I would like to clarify with the p.m.
   a. I would like to get more input from the p.m. as what purpose this product is supposed to serve. Is it for trading purpose? Or is it for risk management or hedging purpose? For the study so far, if the signal has big negative return, the model's predicted return for the ETF could be flat (or slight positive) or negative. At least it is not a 'buy' signal when that happens. Once the purpose of the signal is understood, I could have better judgement whether the signal is good or bad. E.g., the model has RMSE about 1% in the special regime. That 1% accuracy should be viewed in the context of how the signal would be used, i.e., how much tolerance the p.m. can have.
   b. In the previous study, I only tested daily returns. However, the p.m. could have different trading horizon (maybe weeks?), which entails to perform more tests accordingly to see the efficacy of the signal.