Millennium – 4sight Dataset POC

In this report, we provide an assessment on the 4sight dataset by analyzing the data quality and signal effectiveness to predict ETF price. The report contains five segments – data quality check and cleaning, statistical metrics and data assumptions, prediction models, investment strategy back-test, and conclusion and next steps.

<u>Executive Summary</u>

The dataset does seem to contain alphas that could potentially be extracted and leveraged in investment decisions. After cleaning and analyzing the quality of the data, we explore the relationship between signal and ETF price and build rolling regression models to better predict the future stock price movements. While we don't observe substantial improvement in the predictability from the model, a moderate improvement in prediction can help us build a trading strategy that produces superior performances compared to simply holding the ETF. The next step of the project will be understanding the fundamentals of the signal more comprehensively and potentially combining with other signals in-house to better align with our team's investment strategies.

Data Quality Check and Cleaning

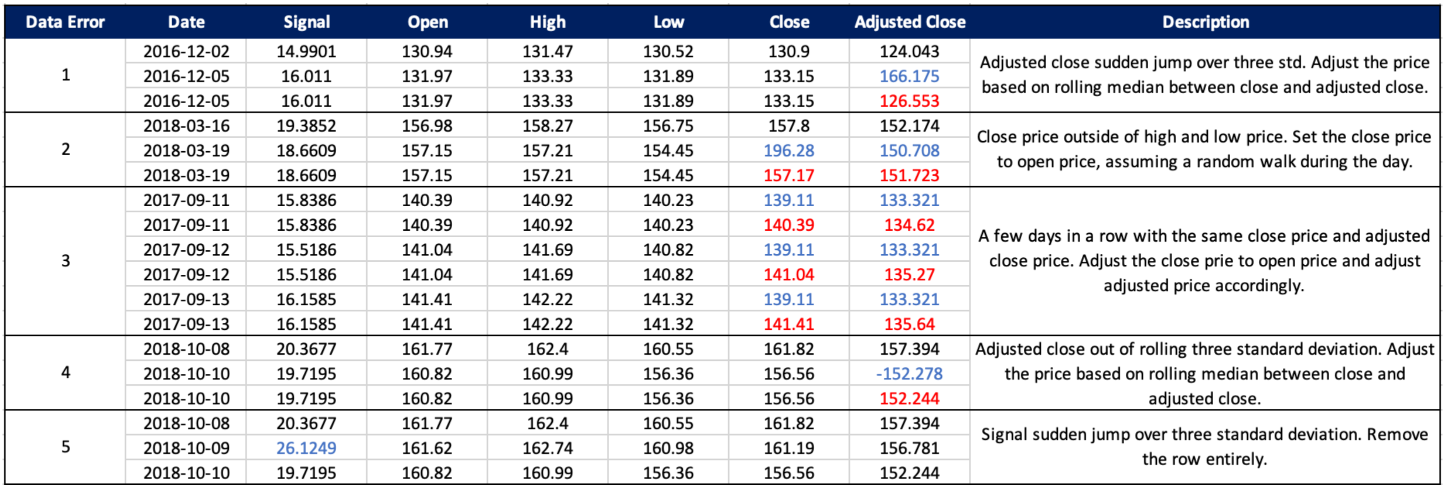
We are given a dataset that includes a signal generated by the vendor along with historical prices for a well-known broad market ETF. The data is collected on daily basis and has a history from November 2015 to January 2020. At first glance, we can instantly uncover some obvious data errors including negative adjusted close price, highest stock price lower than lowest stock price on a given day, some signal values equal to zero at the end of history, and sudden jumps in columns like signal, close price and adjusted price.

We take the following steps to clean up the datasets so that we have high quality data to work with:

- **Validating high/low values**: We ensure that the relationship between column open, column high, and column low holds correctly on a given day by confirming that the column high value is the highest among all the columns, and low value is the lowest among all.
- **Validating close values**: We check that the close price sits between highest and lowest price, and we will set the close price equal to open price if the criteria doesn't hold, assuming fundamental changes of the price have been reflected at the open price and the price movement during the day simply follows a random walk.
- **Validating adjusted close values**: If there're errors in close price, we will adjust the adjusted close price accordingly by subtracting a rolling thirty-days median difference between close price and adjusted close.
- **Validating price outliers**: We make sure that adjusted close price doesn't contain a sudden jump on any given day by calculating a rolling 252 days (one-year trading days) adjusted close price return mean and a rolling 252 days adjusted close price return standard deviation, and update adjusted close prices not within three standard deviations using the rolling thirty-day median method.
- **Validating signal outliers**: We also conduct the three standard deviation tests on column signal, and in this case we remove the days that signals are outliers or signals equal to zero since in the next step we are going to use signal as the independent variable to predict ETF price.

By now, we have completed the data cleaning process and are ready to explore some relationships between the 4sight signal and the dependent variable that we want to estimate – the adjusted close price. We can see some of the data cleaning examples below. (*Note: the number in blue in the below chart are identified as data errors, and the number in red are the adjusted numbers)

Exhibit A



Data Error	Date	Signal	Open	High	Low	Close	Adjusted Close	Description
1	2016-12-02	14.9901	130.94	131.47	130.52	130.9	124.043	Adjusted close sudden jump over three std. Adjust the price based on rolling median between close and adjusted close.
	2016-12-05	16.011	131.97	133.33	131.89	133.15	166.175	
	2016-12-05	16.011	131.97	133.33	131.89	133.15	126.553	
2	2018-03-16	19.3852	156.98	158.27	156.75	157.8	152.174	Close price outside of high and low price. Set the close price to open price, assuming a random walk during the day.
	2018-03-19	18.6609	157.15	157.21	154.45	196.28	150.708	
	2018-03-19	18.6609	157.15	157.21	154.45	157.17	151.723	
3	2017-09-11	15.8386	140.39	140.92	140.23	139.11	133.321	A few days in a row with the same close price and adjusted close price. Adjust the close prie to open price and adjust adjusted price accordingly.
	2017-09-11	15.8386	140.39	140.92	140.23	140.39	134.62	
	2017-09-12	15.5186	141.04	141.69	140.82	139.11	133.321	
	2017-09-12	15.5186	141.04	141.69	140.82	141.04	135.27	
	2017-09-13	16.1585	141.41	142.22	141.32	139.11	133.321	
	2017-09-13	16.1585	141.41	142.22	141.32	141.41	135.64	
4	2018-10-08	20.3677	161.77	162.4	160.55	161.82	157.394	Adjusted close out of rolling three standard deviation. Adjust the price based on rolling median between close and adjusted close.
	2018-10-10	19.7195	160.82	160.99	156.36	156.56	-152.278	
	2018-10-10	19.7195	160.82	160.99	156.36	156.56	152.244	
5	2018-10-08	20.3677	161.77	162.4	160.55	161.82	157.394	Signal sudden jump over three standard deviation. Remove the row entirely.
	2018-10-09	26.1249	161.62	162.74	160.98	161.19	156.781	
	2018-10-10	19.7195	160.82	160.99	156.36	156.56	152.244	

Statistical Metrics and Data Assumptions

Before we start, certain assumptions are made regarding how the signal is generated. **Most importantly, we assume that the signal for point-in-time t is generated and received at point-in-time t-1, and the data does not have fill-delay issues so that a fair comparison between current and historical signal can be made.** With the assumption, we determine that we will be exploring the relationship between signal and price and make predictions at time t + 1 because we will not be able to receive a signal that predicts, for example, time t+30 ETF price. Additionally, we will assign adjusted close price as the independent variable in the model, assuming that signal is generated and reflected all the fundamental changes before point-in-time t and the price movements during time t will not be observed.

With these assumptions, we will be focusing on **three statistical metrics to determine the effectiveness of the signal to predict adjusted close price. The metrics include correlation, direction correctness, and the differences between actual values and predicted values, all on a rolling basis.** We first explore the plain relationship between signal values and adjusted close prices. From the charts below, we can clearly observe that rolling correlation for signal and adjusted close price (purple line in Exhibit C) starts at high 80s but gradually declines and has volatilities at the end of the history. Looking at the metric direction correctness (green line in Exhibit C), which calculates whether the difference between signal at time t and time t-1 is pointing in the same direction as difference between adjusted close price at time t and time t-1, we observe a relatively stable rolling direction correctness at around 50%. A 50% direction correctness is a poor indicator basically telling that the predictability of the signal is somewhat equal to flipping a coin. Further in the analysis, we calculate the rolling correlation on signal return and adjusted close return to identify the effectiveness of signal to predict actual return. From the chart below we can observe that return correlation fluctuates around zero at the beginning of the history, and while the correlation increases a bit during certain periods in 2018 and 2019, it doesn't really stabilize and start to decline after mid-2019.

In summary, we couldn't reach the conclusion that using the plain signal is promising to predict the price movement of the ETF. We will continue the analysis and build prediction models to see whether we can uncover more hidden relationships between signal and stock price and use the signal to build an effective trading strategy.

Exhibit B

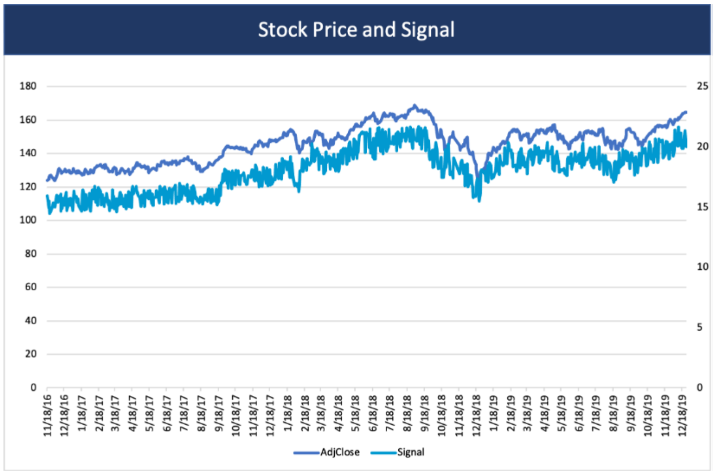
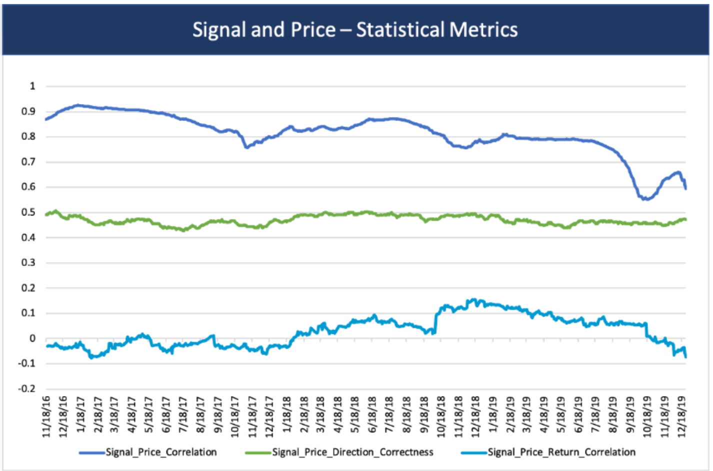


Exhibit C



Prediction Models

In this section, we will discuss the prediction models we built to uncover the relationship between signal and stock price. Since the goal of the study is to analyze whether alphas can be extracted from the signal, we're assuming that the signal has been designed to predict the stock price and thus we fit regression models to further uncover any hidden values that could not be extracted by simply using plain signal values.

We first explore whether fitting an OLS model to the signal can better predict adjusted stock price. We build a rolling 252 trading days OLS model given any point-in-time t and use the model to predict stock price at time t + 1:

- **Examining price correlation:** Looking at the rolling correlation for prices in Exhibit E (purple line at the top), we observe that correlation steadily declines from a high-80s percentage at the beginning of history to a mid-50s percentage with some volatility at the end of the history.
- **Examining model r-squared:** To understand how well the stock price movement can be explained by signal movement, we look at the in-sample r-squared and out-sample r-squared. From the light blue and gray lines in Exhibit E, we can see both lines follow a similar pattern as the rolling correlation, with in-sample r-squared starting at high-60s percentage and declining to low-40s percentage and out-sample r squared starting at low-fifty percentage and declining to low-twenty percentage. We can also easily observe the fact that the model is not doing as well in out-sample data as in in-sample data.
- **Examining direction correctness and return:** Lastly, we take a look at rolling direction correctness and rolling correlation of stock return and predicted return. From the green line, we can observe that direction correctness is staying fairly stable at fifty to sixty percent through the history, which is slightly better than using the plain signal values. From the blue line at the bottom of the chart, we see that correlation on returns stay roughly between twenty percent to thirty percent, which also improves a bit from simply taking the return from the signal.

In summary, we do see some value been extracted by leveraging the OLS model. However, from the metrics we learnt that the out-sample performance is lower than in-sample performance, which might suggest we are overfitting the data. In order to make the model more generalized, in the next step we are going to fit a regularized regression model, in hopes to reduce the out-sample variance by introducing some bias into the model.

Exhibit D

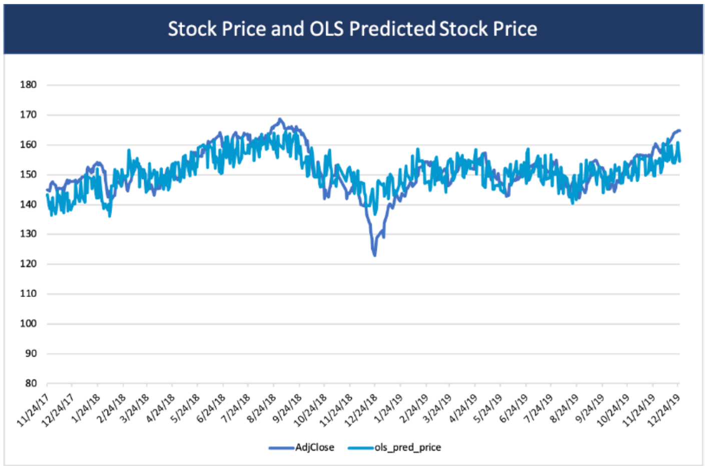
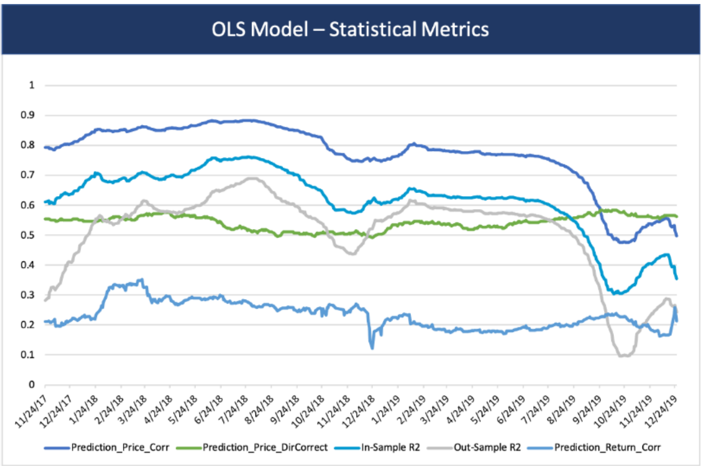


Exhibit E



As discussed above, we then explore fitting a regularized regression model to see if we can better predict adjusted stock price by desensitizing the signal and introducing some bias into the model. We will leverage Ridge regression model for the analysis. Using a very similar concept, we fit a Ridge regression model to the data on a rolling 252 trading days basis and predict the stock price at time t+1.

In order to find the best lambda value to use in the loss function, we leverage GridSearchCV function in the Sklearn package and randomly split our training data into training sets and testing set. Looking at the statistical metrics below, we can instantly find that the metrics look very similar to the OLS regression results. While in-sample r-squared decreases slightly, we don't see an improvement in out-sample r-squared. In addition, all the fluctuations and the sudden drop at the end of history are not solved by leveraging the regularized model, and we also don't observe obvious improvements in terms of other metrics like correlation and direction correctness.

In conclusion, we believe that introducing a regularization is not going to decrease the variance in out-sample or improve model performance. We will be using a simple rolling OLS model as our primary model and do an investment back-test in the next step to see what performance we will generate if we adopt this model to build a trading strategy.

Exhibit F

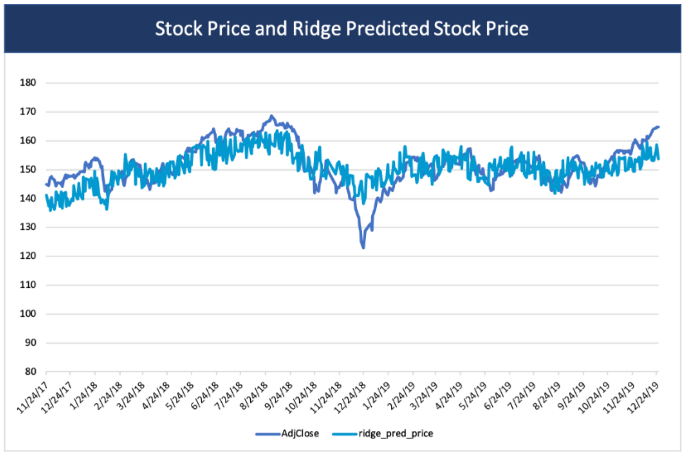
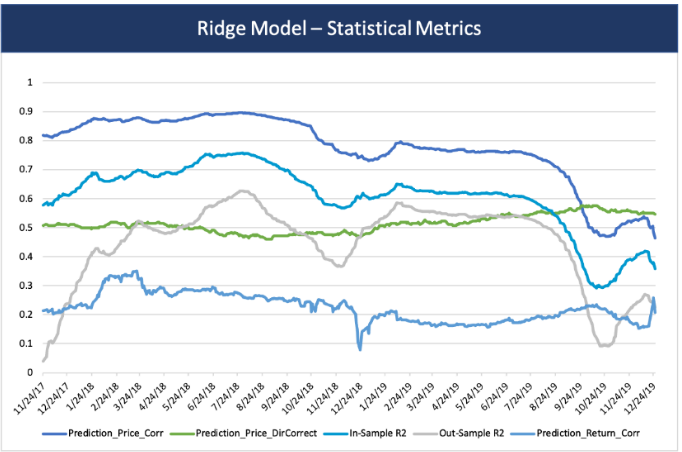


Exhibit G



Investment Strategy Back-Test

In this section, we are building an investment strategy based on the OLS regression model that we built. We will establish a simple trading strategy that if the prediction is positive for time t, we will long the ETF, and if the prediction is negative for time t, we will short the ETF. We will compare the performance between this strategy and a simple long strategy that holds the ETF for the entire history no matter what the model is predicting. Additionally, to simplify the exercise we will assume there's no trading cost for all the transactions.

Even though we don't have an extremely promising model with perfect predicting accuracy from above, we can observe that back-test result is meaningful. From Exhibit H, the green line is the portfolio performance assuming we are solely making transactions based on the model prediction. **The portfolio based on our predicting model provides a 250%+ growth from year 2016-19, while the simple hold strategy only provides a growth of ~30% over the same timeframe.** We continue our analysis on the stability of each portfolio return. In Exhibit I, the green line is the rolling 252-day return in the model portfolio, the blue line is the rolling 252-day return in the base portfolio, and the gray and yellow lines represent the volatility in each of the rolling return. We can easily observe from the green and gray line, which represent model portfolio return and return standard deviation, that **the model portfolio's return is slightly more volatile compared to base portfolio, but it provides a much richer return profile.**

In summary, while we don't observe huge difference in volatility, we do see a huge difference between the two portfolio rolling returns. Model portfolio's return constantly generate a 40%+ return on a rolling 252-day basis and even generate 60%+ returns in the near term, while the hold-strategy's performance struggles in the near term as the ETF is not performing well.

Exhibit H

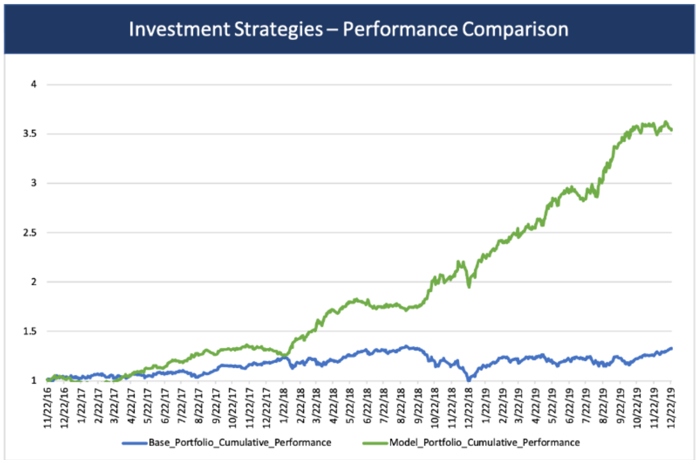
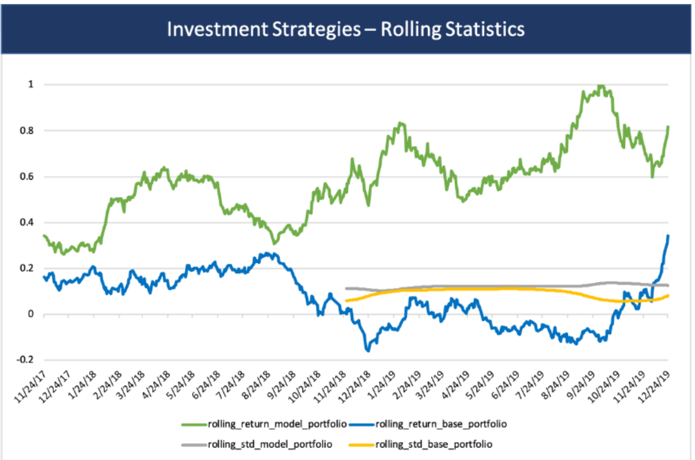


Exhibit I



Conclusion and Next Steps

We have completed several steps to assess whether the signal from the vendor can be useful to predict the ETF price and potentially build a trading strategy. From the previous discussions and analyses, we conclude that **the signal does contain alpha and can improve the prediction of stock price movements**. Even though the improvement is not substantial, with a rolling return correlation around 20% and a rolling direction correctness slightly above 55%, we still found a potential alpha that we can leverage when building our investment strategy and back-testing it. One obvious reason is that by leveraging the signal from the vendor, we are not making predictions on factors that move the stock price, for instance, top-line revenue, but we are making predictions directly on the stock price movement. Because of that, **even with a moderate improvement on the predictability, we are observing huge value that can be extracted from the signal**.

Although that we are observing promising results from the back-test, we believe that this POC is far from complete and at least three projects that could help better leverage the signal and reach a stronger conclusion on the POC.

We list the next steps as the following:

- **Fundamental understanding of the signal**: After receiving the signal, we did a **correlation test with the ETFs on the market and try to identify the ETF in question.** After this preliminary analysis, **we found high correlations between this ETF and broad US market stock ETFs, such as VTI and IWM.** Realizing that the signal is an indicator of the broad US market stock, we want to understand more about how the signal is generated. For instance, what are the datasets and variables that are been used by the vendor to produce the signal and whether we believe those datasets and signals are comprehensive enough to produce it, who are the vendors to those underlying signals and whether there will be fundamental changes to those signals, and what are the methodologies including how the datasets are normalized or dealt with the fill-delay issue and the methodologies to construct the final signal provided to us. Even though we believe there may be IP roadblocks that could potentially prevent the knowledge from being shared to us, we believe we need more clarification and would lack too much fundamental knowledge otherwise.
- **Combination with other signals in-house**: We surely own other datasets and signals in-house. In the previous exercise we're using this 4sight signal alone to predict stock price and we can observe some alphas extracted. If we can combine this particular signal with other indicators we have in house, we could potentially build a more sophisticated model and make the data tell a more comprehensive story. For example, if we have some kind of credit card data, website data, or hospital claims data, we could potentially generate some market monitoring signals from these datasets and join with the 4sight signal to further enhance our belief on what the signals are telling us.
- **Alignment with trading strategies**: In this exercise we have an implicit assumption that the investment team will want to trade this broad US market ETF. Although that assumption is highly likely to hold, we still want to make sure that alphas we extract can actually be used in the investment process. Further, this signal might not only be used to predict this particular ETF. Assuming there could be some bias in the signal, for instance if a lot of the variables that are used to generate the 4sight signal are macro and credit card related, then we might want to further analyze whether the signal can even better predict a subset of stock names or certain sectors like consumer sector.
- **Improvement of the existing models**: Even though we have come to a stopping point with the prediction models above, we believe there're still improvements to make on those models. For instance, we see a decline in model accuracy at the end of the history, however, the direction correctness is not as impacted and thus the portfolio continues to maintain a high performance. This is absolutely something that we can look into further in the next phases of this project.