Data Cleaning of Stock Market Data and Price Forecasting

Stefanos Baros

February, 18, 2021

1 Data quality checks

1.1 Code files

This report relies on the python code file **Data-Cleaning-Stock-Market-Data-Stefanos-Baros.py**. The cleaned data set is exported into a file named **clean-dataset.csv**.

1.2 Consistency errors

By observing the structure of the dataset, we identified several consistency checks we needed to carry out in order to be certain that the quality of the data was high. For the given dataset, four types of consistency errors seemed appropriate:

- **Error type 1**: when low price is not the lowest one
- **Error type 2**: when high price is not the highest one
- **Error type 3**: when adjusted price is higher than the closing price
- **Error type 4**: when any price is negative

The following rows (described by their indices) were identified as having errors belonging to the above categories.

- **Error type 1** = {407, 455, 456, 457, 458 ,459, 460 ,461, 462, 463, 464, 577, 671, 739, 892, 966}
- **Error type 2** = {431, 585, 766, 983}
- **Error type 3** = {262, 338}
- **Error type 4** = {732}

There were 16 rows with errors of type 1 as described above. Among those, 10 were consecutive rows and the remaining 6 were other isolated rows. We corrected the values in the 6 isolated rows using the *last observation carried forward method* (LOCF) method of imputation. For the 10 consecutive rows, LOCF was not a suitable approach. To correct the values in those rows, we carried out linear regression 5 times by excluding these rows from the process. We obtained 5 models where each one of those had the signal as input feature and as output the low, high, open, close and adjusted close price, respectively. After verifying the performance of these 5 models using *goodness-of-fit* and *prediction error metrics* (discussed later) we used the Signal values in these 10 consecutive rows to predict all prices and filled out the corresponding columns.

The rows with errors of type 2 and type 3 were isolated therefore once again we used the *last observation carried forward method* (LOCF) of imputation to correct their values. Basically, in each isolated row, we corrected the prices using the data from the previous row.

There was only one row with *error type 3* having a negative adjusted closing price. The absolute value of this price seemed consistent with the other prices in that row therefore we assumed that this was a typo. We corrected this value by simply replacing the adjusted closing price with its absolute value.

1.3 Duplicates

We also scanned the dataset for dublicated rows and did not find any.

1.4 Outliers

We first plotted all prices over time to see whether there were any outliers. Outliers were spotted in the Signal plot. To formally identify the outliers in our dataset we performed a *statistical test using the median absolute deviation around the median criterion (MAD)* that relies on the following metric:

$$score = \frac{|\text{Signal} - \text{median}(\text{Signal})|}{\text{MAD}} \tag{1}$$

Given that the Median Absolute Deviation (MAD) is an approximation of the standard deviation we chose the threshold *to be 3*, which is standard for this statistical test. The following outliers were identified (described by their indices):

- **Outliers** = {500, 731, 1032, 1033, 1034, 1035, 1036, 1037}

The last 6 corresponded to the observations that had Signal value zero. We assumed that those observations were not actual outliers but instead had *missing values.* The observations 500 and 731 were eliminated from the dataset as there was no way for us to correct their values in a systematic way.

1.5 Missing values

To correct the zero values, which we assumed were missing values, in the Signal column of the last 6 observations in our dataset we developed a *Holts-Winters*

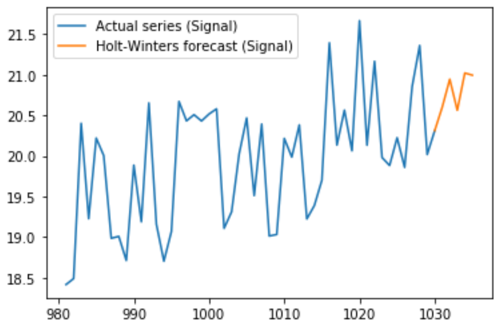

Figure 1: Holts-Winters triple exponential smoothing time-series forecasting model

triple exponential smoothing time-series forecasting model using all observations except those. The seasonality period was chosen to be 260 corresponding to the total days in a year excluding weekends where price data are usually available. The following figure depicts 50 observations and the forecasted 6 ones obtained from the *Holts-Winters triple exponential smoothing time-series forecasting model*.

2 Effectiveness of Signal in predicting ETF price

To assess the effectiveness of Signal as predictor of the different prices we used the correlation matrix and linear regression analysis.

2.1 Correlation matrix

First, we computed the correlation matrix which describes the correlation between the Signal and the different price values. We immediately realized that the Signal is *very highly* correlated with all price levels (low, high, open, close, adjusted close) which implies that the Signal is a *very good, unrealistically one could argue, predictor* for those prices. We also plotted the Signal versus the different prices to get a feeling of the general shape of these characteristics.

2.2 Linear regression

To formally verify whether the Signal is a good predictor of the ETF price, we performed linear regression. We performed linear regression repeatedly, using every time the Signal as an input and as an output either the low, high, open, close or adjusted close prices. We trained the regression model on 75% of our data and used the remaining 25% for validation. The performance of one of the five models, corresponding to the adjusted closing price, is shown in the figure below.

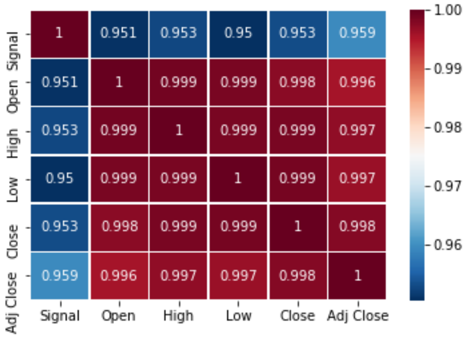

Figure 2: Correlation matrix, Signal vs prices.

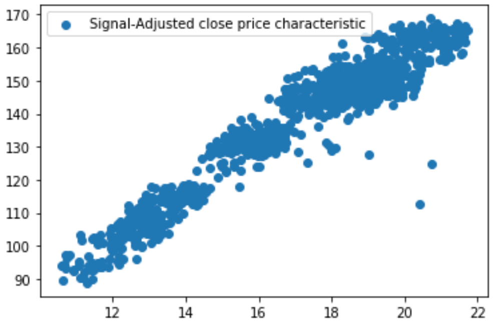

Figure 3: Signal - Adjusted closing price characteristic

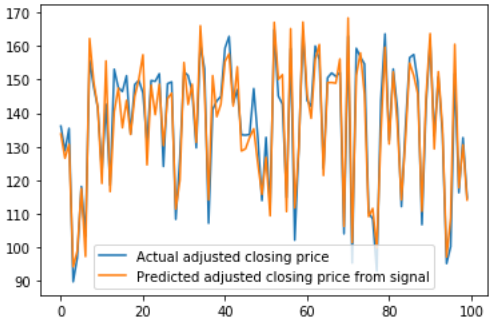

Figure 4: Regression model prediction, adjusted closing price.

2.3 Goodness-of-fit criteria and prediction error metrics

To formally assess the quality of our models we used both a goodness-of-fit criterion and prediction error metrics. To assess the goodness-of-fit, we computed the *adjusted R squared* value. For the model corresponding to the adjusted closing prices we obtained:

$$\text{Adjusted } R^2 = 0.945 \tag{2}$$

We also computed the *percentage mean squared* and *percentage mean absolute errors*:

$$MSE = 1.397\% \tag{3}$$
$$MAE = 0.013\% \tag{4}$$

Similar values were obtained for the remaining 4 regression models. The adjusted R squared value is *close to 1 and the MSE and MAE are both extremely low.* From these metrics, it appears that the Signal is *an unrealisticallly good predictor* of the different price levels. If the Signal is indeed such a good predictor of the market prices as this preliminary analysis shows, then this Signal can enable us to make *very high* profits from the market.

3 Recommendations to the Portfolio Manager

Overall, the dataset had 6 missing values and 2 outliers. Besides that, the Signal appears to be *an unrealistically good* predictor of the market prices. This together with the fact that we do not know exactly how much future information was really incorporated in the process of generating the Signal values, I would recommend that we are cautious and only buy a few datasets for a couple of months, try them out see how much profit they can earn us and proceed to buy more data sets after we are totally convinced that the Signal is indeed such an overly good predictor of the market prices as it appears from our preliminary analysis.