Application of Statistical Bootstrapping in Financial Machine Learning

James Hopham

August 2, 2024

Financial data is notoriously difficult to work with due to the many statistical properties it violates that are generally required to produce a functioning model, such as stationarity and independence, requiring us to use various techniques to make reasonable decisions to work with the data. The objective of this project was to employ techniques against unstructured financial time-series data by applying approaches in mathematical and statistical modeling thereby improving key properties necessary for robust and reliable models. This project implements a few of the concepts covered by Marcos Lopez de Prado's textbook Advances in Financial Machine Learning including ideal financial data structures for improved statistical properties, differentiation, data labeling and sampling, sequential bootstrapping, and bagging machine learning models. In this project I employ these techniques to create an ETL pipeline to a random forest model for E-mini S&P 500 futures minute data from 2009 to 2018 according to classifications determined by the triple-barrier method.

Introduction

Financial market data is plagued with many issues including information leakage, multiple testing bias, and non-IID distributions that make them suboptimal for mathematical analysis and more difficult than working with cleaner data in other industries such as the biomedical industry. In addition to this, financial data is highly correlated to past information which requires us to consider trade-offs in stationarity and memory. In this project, I focus in particular on the characteristics of autocorrelation, stationarity, independence and variance through the discussions and approaches detailed in the textbook *Advances in Financial Machine Learning* by Marcos Lopez de Prado where he addresses many issues and dilemmas faced by practitioners in finance.

Quantitative analysis of unstructured financial data requires us to extract significant information and store it in a regularized format such as bars to generate a continuous, homogenous, and structured dataset. Standard time bars such as those presented in stock market data vendors are generally suboptimal due to a multitude of reasons; most notably, they exhibit poor statistical properties such as heteroscedasticity and autocorrelation. These extend into further issues due to the lack of contiguous data leading to methods such as problems in sampling due to additional concerns in reliability as a result of overlapping information among individual samples. Machine learning algorithms that do not scale well with sample size require sampling techniques that allow the model to be exposed to the most highly relevant instances of

the data. To correct for this, alternative transformations such as volume bars can allow us to obtain improved statistical properties in the time series data.

To address autocorrelation, Marcos Lopez de Prado suggests using fractional differentiation as an alternative to standard integer differentiation. This approach attempts to account for the memory-less returns from the Box-Jenkins approach and the non-stationary returns from the Engle-Granger approach. Marcos Lopez de Prado expands upon the original fractional differentiation implementation introduced by Hosking [1981] by suggesting using a fixed-width window to drop weights after their modulus ($|\omega_k|$) falls below a given threshold value (τ). The goal is to compute the minimum d such that the p-value of the ADF test on FFD(d) falls below 5%.

To filter financial time series for significant data, event-based sampling can be done to extract important signals from the data based on various measures such as macroeconomic statistics or abnormal changes in volatility. Marcos Lopez de Prado [2018] presents the cumulative sum filter as one example of a useful event-based sampling method where a bar t is sampled if the cumulative sum $S_t \geq h$ where h is the threshold for the filter. When considering observations $\{y_t\}_{t=1,\ldots,T}$ where $S_t = max\{0, S_{t-1} + y_t - E_{t-1}[y_t]\}$, the filter is triggered if there exists a time τ within the interval [1,t] when $S_t \geq h \Leftrightarrow \exists\tau \in [1,t] \mid \sum_{i=\tau}^{t} (y_i - E_{i-1}[y_t]) \geq h$. This method prevents insignificant price movements from being sampled and allows us to only extract meaningful time series fluctuations.

In order for us to apply supervised learning models to our financial time series, our data must be labeled to predict desirable outcomes of interest. The triple-barrier method can be used to label observations based on dynamic thresholds computed by daily volatility to avoid situations where $\tau >> \sigma_{t\,(i,\,0)}$ or $\tau << \sigma_{t\,(i,\,0)}$. An observation is assigned label $y_i \in \{-1, 0, 1\}$ according to the value of the price return $r_{t_{i,0}, t_{i,0}+h}$ relative to τ. The method takes into account the entire path $[t_{i,\,0}, t_{i,\,0}+h]$, where h defines the vertical barrier. Once the labels are assigned, the data can be examined for extremely rare labels and investigated to determine whether these should be dropped to avoid bias in the model.

Despite having labeled data, we must address the issue of non-IID observations when sampling. Specifically, $\{y_i\}_{i=1,\ldots,I}$ are not IID when $\exists i \mid t_{i,1} > t_{i+1,\,0}$. Since imposing too many time horizon restrictions risks limitations on the model, we assign weights that allow some overlap but restrict based on uniqueness to correct for influences from overlapping outcomes. To address this, we identify concurrent labels at time t, $c_t = \sum_{i=1}^{I} I_{t,i}$, where $I_{t,i} \in \{0,1\}$ with respect to

whether $[t_{i,0}, t_{i,1}]$ overlaps with [t-1,t]. A label's uniqueness is $u_{t,i}=l_{t,i}c_t^{-1}$ and the average uniqueness of label i over its lifespan is $\bar{u}_i = (\sum_{t=1}^{T} u_{t,i})(\sum_{t=1}^{T} l_{t,i})^{-1}$.

Although bagging is a useful technique for improving predictions, we must make adjustments when bootstrapping to account for redundant information by using the average uniqueness of our labels. We cannot incorrectly assume our draws are IID and can use sequential bootstrapping where subsequent draws are made based on updated probabilities $\delta_j^{(2)}=\bar{u}_j^{(2)}(\sum_{k=1}^{I} \bar{u}_k^{(2)})^{-1}$ where $\{\delta_j^{(2)}\}_{j=1,\ldots,I}$ are scaled up to 1, $\sum_{j=1}^{I} \delta_j^{(2)}=1$. This allows us to generate more diverse samples that are less biased by overlapping events by decreasing the probability of drawing repetitive information. We can then weight our bootstrapped samples to account for highly overlapping outcomes where $\tilde{w}_i=|\sum_{t=t(i,0)}^{t(i,1)} \frac{r(t-1,t)}{c(t)}|$ and normalized as $w_i= \tilde{w}_i I(\sum_{j=1}^{I} \tilde{w}_j)^{-1}$.

Methodology

In this project first I begin the ETL pipeline by extracting data from unstructured E-mini S&P 500 futures minute data from 2009 to 2018. I first check and remove outliers and follow with a transformation of the data into dollar bars and then check for autocorrelation in the returns of the dollar bars. The transformed dollar bars are sampled to only observe for significant events by using the cumulative sum filter that consists of a threshold that considers the daily volatility of the time series' price movement. I then label the events with the triple barrier method to identify profit taking and stop loss barriers by dynamically considering volatility and additionally pass a vertical barrier representing a limit to the holding period.

I then begin an implementation of meta-labeling by using bollinger bands to identify the side of a bet, then train an ML model to label binary outcomes to indicate whether we should either pass on or take a bet. The meta-labels and the ML model are separate from the primary data we began with. Meta-labeling should help us increase our F1 score and reduce FPs. We are willing to sacrifice precision for better recall in this case. Since poor results were obtained upon training a random forest model, I repeat the experiment by starting with fixed-width window fractional differentiation to account for serial correlation and repeat the steps detailed previously.

Upon obtaining our transformed data, I calculate the number of concurrent labels and compute the average uniqueness values to construct an indicator matrix to be used for further sampling. The matrix is used to employ sequential bootstrapping to consider overlapping events. These bootstrapped events are then fed to a random forest model.

Random forest corrects for the overfitting problem faced by decision trees by training multiple decision trees on bootstrapped data to produce ensemble forecasts. Unlike standard bagging, random forests additionally adds randomness by subsampling a set of features at each node split to further decorrelate the estimators. As detailed previously, this methodology compels us to take further considerations against non-IID data to avoid overfitting. We integrate sequential bootstrapping to replace the standard bootstrapping that would be used for random forest.

Results

Figure 1. Fractional Differentiation Test

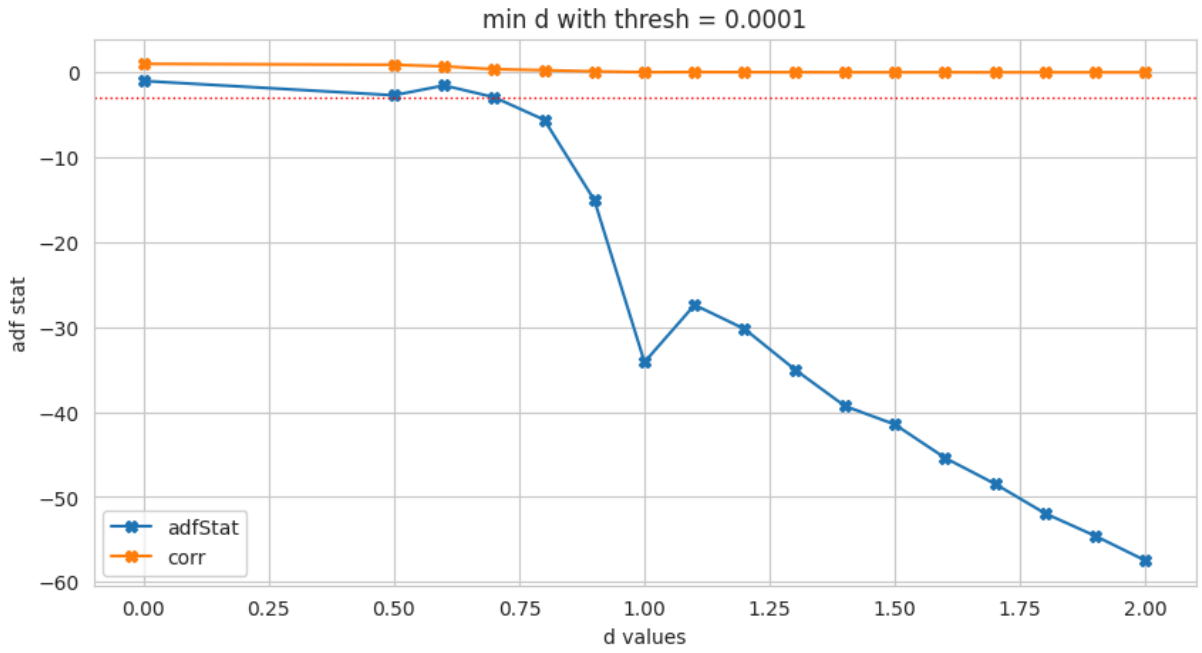


Figure 2. Test Metrics

```
Accuracy: 0.903337169159954
Classification report:
              precision    recall  f1-score   support

          -1       0.79      0.42      0.55        79
           0       0.92      0.98      0.95       765
           1       0.30      0.12      0.17        25

    accuracy                           0.90       869
   macro avg       0.67      0.51      0.55       869
weighted avg       0.89      0.90      0.89       869
```

Figure 3. Confusion Matrix

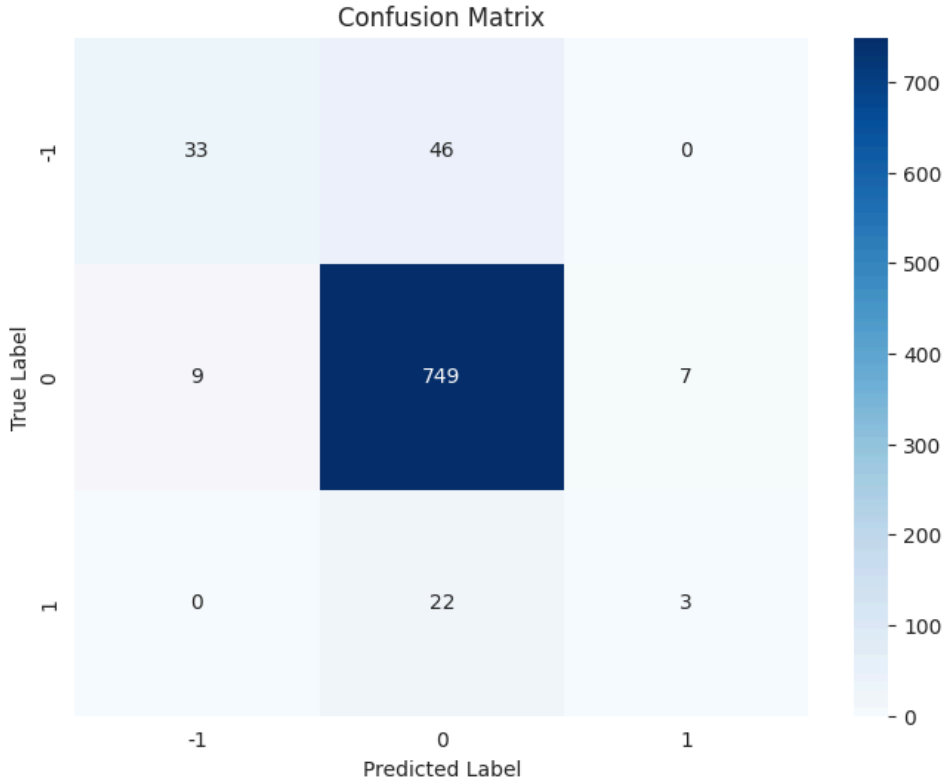
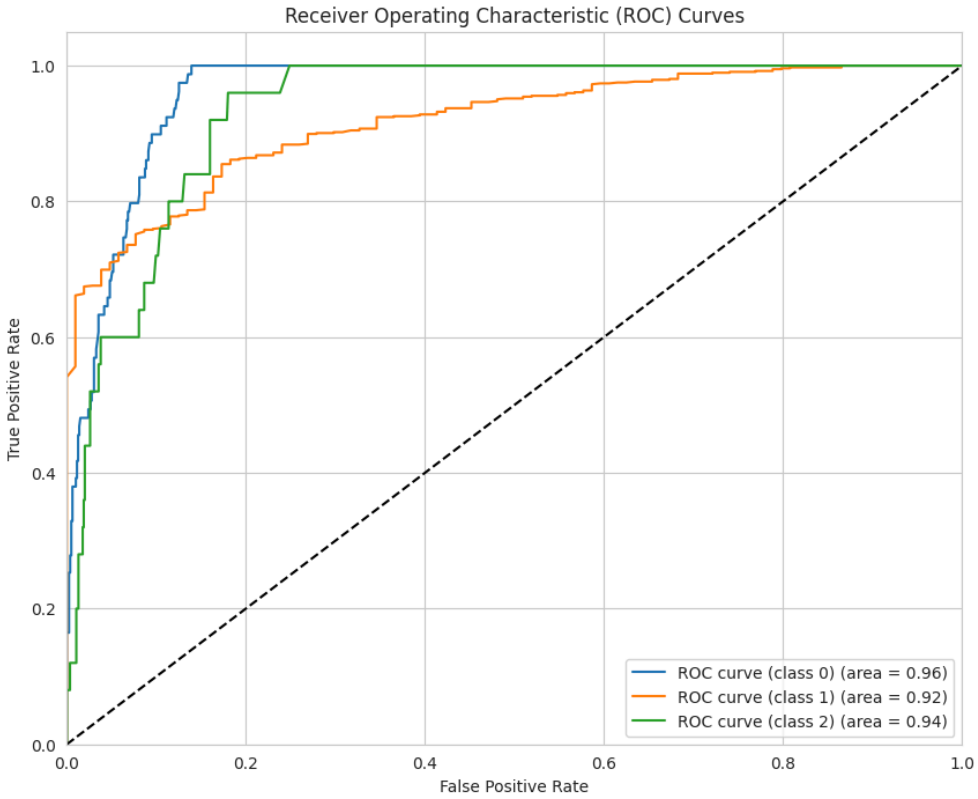


Figure 4. ROC Curve



Further Discussion

The test demonstrates positive results; however, further considerations should be taken to correct for multiple testing biases. Additionally, hyper-parameter tuning in cross validation should be further scrutinized to ensure leakage in the data does not occur. I plan to implement these in future projects.

References

Hosking, J. R. M. (1981). Fractional differencing. *Biometrika, 68*(1), 165-176.
López de Prado, M. (2018). *Advances in Financial Machine Learning*. John Wiley & Sons.