**Backtesting.**

Conventional statistics traditionally assume that observations are independent and identically distributed (IID).

However, financial data has an obvious correlation when sampling data adjacent in time \cite{shumway2000time}.

This type of data, where the temporal order of observations matters, is known as time series.

With time series, the traditional ways of analyzing data can't be used \cite{shumway2000time}.

Conventional Cross-Validation ignores the temporal component of forecasting using time series and will therefore not work with this problem \cite{de2018advances}.

Instead, backtesting is used when evaluating the performance of a time series forecasting model.

Backtesting is a historical simulation of how the model would have performed should it have been run over a past period \cite{de2018advances}.

Essentially, past performance on historical data implies future performance on new data that has not yet been seen.

To give an accurate reflection of the performance of the model, real-world conditions must be emulated as closely as possible.

Transaction costs and other fees need to be accounted for.

Backtests, however, will never be completely accurate, as the impact of the strategy on the market cannot be modeled.

Backtesting attempts to mimic real-world conditions by training on data up to some point in time, and then making predictions of what's to come next \cite{de2018advances}.

The model only has access to data up to what is present for the model and must make predictions on data it has not seen yet.

After making predictions, the dataset the model has access to increases sequentially to include the period it just made predictions on.

This continues until the end of the dataset is reached.

There are some variations of backtesting.

The model can be trained once or be refit for every iteration.

The training-set size can be expanding or a sliding window as the model progresses through time.

For high-frequency data such as hourly time series sliding window might be more appropriate as it has a faster training time.

For longer time intervals it might make sense to utilize the whole data set as the total observations are limited.

--- Figure illustrating backtesting.

Sliding constant size training set.

Expanding training set.

Without refit

The signal-to-noise ratio of financial data is low \cite{taleb1997dynamic}.

Exchanges have only traded oil futures for about 40 years.

With limited data, training and testing the model is a challenge.

The risk of modeling random noise is substantial, leading to false positives, i.e., models that generalize poorly.

False positives should be minimized in a way that does not hinder our ability to discover true positives \cite{arnott2019backtesting}.

Backtesting should emulate the scientific method where a hypothesis is developed, and empirical research is conducted to find evidence of inconsistencies with the hypothesis.

It should not be confused as a research tool for discovering predictive signals.

It should only be conducted after research has been done. It is not an experiment, but rather a simulation that give us a chance to see if the model behaves as expected \cite{de2018advances}.

Random historical patterns might exhibit excellent performance in a backtest.

However, it should be viewed with caution if there is no ex-ante logical foundation to explain the performance \cite{arnott2019backtesting}.

Furthermore, only live trading can be considered truly out-of-sample.

Having experienced the hold-out sample provides insight into what made markets rise and fall, a luxury not available until after the fact.

When conducting tests, it is imperative to keep track of how many strategies are being examined.

One of twenty random strategies would likely exceed the two-sigma threshold by pure chance.

Therefore, the threshold should be higher when numerous uncorrelated techniques are attempted \cite{arnott2019backtesting}.