Backtest Results in an Overfitting Model
Considerations about its application in the final project

Submitted to:
Michael Rolleigh
in partial fulfillment of the course
DAT 5308
Algorithmic Trading in R/Python

Submitted by:

Jorge Andrés Betancourt

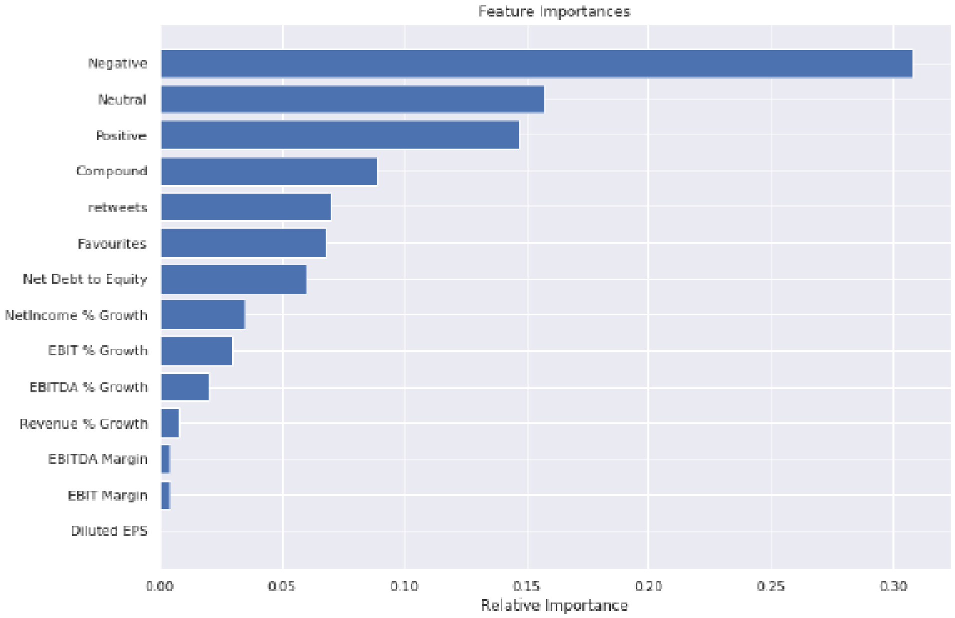
Hult International Business School
Boston, Massachusetts
Date: June 20, 2019

This project had as objective to prove the results of an overfitting model using fundamental and alternative data within five years (2014-2019) for RedHat cloud computing company. The data is formed by revenue percentage growth, EBIT percentage growth, number of retweets, among others. Machine Learning performed an essential role in this analysis, being used to predict returns based on the data. To develop a more robust algorithm, this paper will discuss a different approach to improve the model based on the results.

Data Collection

For the data collection process to perform a text/sentiment analysis, we used the twitter API to collect information about RedHat. For the fundamental data collection, Yahoo Finance served as a source. Considering data security aspects, this data that was used as an input in the model has been publicly available for the period of the analysis. It is also essential to take into consideration the periodicity of the data that is used, as it can change the results. Therefore, the fundamental data were collected quarterly for this project.

To have a better understanding of the importance of the data for the analysis, exhibit 1 describes all the variables that were used in the model, in this case, the cumulative number of negative tweets have higher relative importance. Therefore, fundamental data could be considered as less important than alternative data. In consideration, for the development of the model, we assume an increment of the alternative data granularity by assigning weights to important events (e.g., IBM acquisition of RedHat).



Cross Validation

Accordingly, to Prado (2018), we have to take into consideration that: "one reason k-fold cross validation fails in finance is that observation cannot be assumed to be drawn from an independent and identical distributed random variables process. The second reason for cross-validation failure is that the testing set is used multiple times in the process of developing a model, leading to multiple testing and selection bias. Leakage takes place when the training set contains information that also appears in the testing set."

The initial machine learning model has been developed bases on a decision tree without constrains in its max depth to exemplify the easiness and danger of overfitting. To demonstrate the overfitting degrees, 5-fold cross-validation was performed.

To correct this problem, Prado (2018) recommends, first, purging serially correlated observations from the training set and second eliminate observations from the training set that immediately are followed by an observation in the test set.

Analysis

Using the bt.packaged, we defined three strategies for the trading algorithm, first rebalance according to the target weight (TW), as the signal from the decision tree (Predicted Returns): +2 Strong buy when the TW is +1.0; +1% Medium Buy when the TW is +0.5 and in the opposite side -1% Medium Sell when the TW is -0.5; -2% Strong sell when the TW is -1.0.

The second strategy is modifying the results by adding limit deltas to limit the change in portfolio no more than 10% per day by using bt.algos.LimitDeltas().

The third strategy is adding limit deltas and one day lag since there is no way to know the closing price before the market close.

What we conclude with our study is when financial analysis does not control for overfitting past performance is not a good indicator of future performance. One way to prevent backtest overfitting for future analysis is to compute performance degradation and the probability of loss to get a more realistic result. Finally, it is necessary to implement grid-search with purged k-fold cross-validation to develop a more robust machine learning model

1. References

Prado, M (2018). *Advances in Financial Machine Learning.* United States, New Jersey: John Wiley & Son,Inc.