Financial Machine learning

Considerations about its application in the final project

https://github.com/dieko95/AlgoTrading

Submitted to: Michael Rolleigh
Submitted by: Diego Gimenez
Submitted at: 06/20/2019

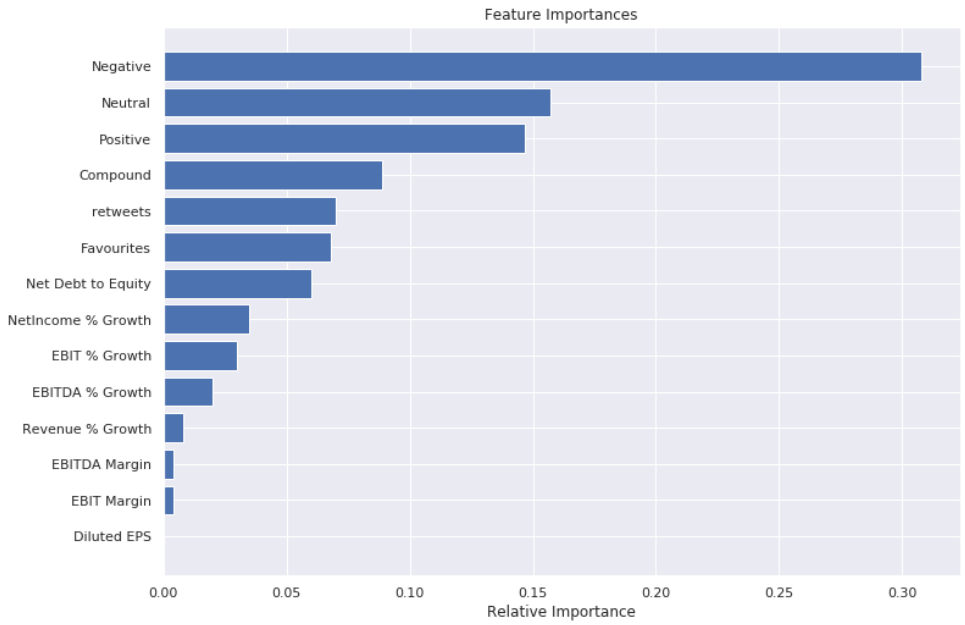
Machine learning had an important role in our final project. Its purpose was to predict returns based on alternative and fundamental data. These two types of data include sentiment analysis, number of retweets, number of favorites, revenue percentage growth, EBIT percentage growth, among others. The following paper will discuss what the team could have done differently in order to develop a more robust algorithm to extract the signal from alternative and fundamental data.

Data Collection

Prado (2018) stated “Backtesting is not a research tool. Feature importance is.” (p. 114). Following this train of thought, mean-variance reducti on was implemented with Scikitlearn’s random forest regressor in order to examine relative features importance.

The data that was used as an input in the model has been publicly available for a considerable amount of time. Twitter has its own API, easily accessible by anyone that has enough patience to watch YouTube tutorials. Moreover, fundamental data can be easily accessed through Yahoo Finance. Nonetheless, it is necessary to emphasize the importance of alternative data in the final project’s machine learning model.

As seen in the graph below, the number of negative tweets had higher relative importance and alternative data shows to be more important than fundamental data. Thus, for the future development of the model, it is plausible to hypothesize an increment of the alternative data granularity. This increment of detail can be done by assigning weights to important events (e.g. IBM acquisition of Redhat) or monitoring the competitors’ tweets. Finally, it is highly important to do further research on how neutral tweets may bias the model. In average, 19 neutral tweets were identified per day. However, 0.54 negative tweets were identified per day.



Cross-Validation

The first machine learning model that was developed, corresponded to a decision tree without constraints in its max depth. This was done with the purpose of exemplifying the easiness and subsequent danger of overfitting. Moreover, 5-fold cross-validation was performed in order to illustrate how overfit is the model. Prado (2018), states why k-fold cross validation fails in finance. Standard cross validation assumes that variables follow an independent and identically distributed process (IID). However, trading observations do not follow the same probability distribution and are not mutually independent. This leads to data leakage between the train and test sets.

In the further development of the machine learning model, it is necessary to consider two plausible solutions recommended by Prado (2018). First, purging serially correlated observations from the training set. Second, eliminating observations from the training set that immediately are followed by an observation in the test set.

Hyperparameter Tuning

Hyperparameter tuning was developed by performing a 5-fold cross-validated grid-search. The optimum depth was 3 and the optimum max number of leaves was 5. The overfitting was reduced considerably. The R squared of the training set and test set decreased from 0.99 to 0.03 and from -0.74 to -0.26. Due to the limitations explained in the latter section (non-IID observations and data leakage), it is necessary to implement grid-search with purged k-fold cross-validation in order to develop a more robust machine learning model.

References

Prado, M. L. (2018). *Advances in Financial Machine Learning*. Wiley.