Forecasting Financial Data Using Ensemble Learning and Technical Indicators

Project Plan

Abhinav Goyal (3035240130)

29th September 2019

Supervisor : Dr. Yip Beta

1. Project Background

Over the past decade, there have been a plethora of studies conducted in the finance industry exploring the use of machine learning to predict financial data such as stock prices. However, according to [1] a majority of these studies tend to focus on the application of a single machine learning model on a specific asset class. Training a single model on a single market dataset can lead to an inadvertent bias and overfitting, it can also adversely affect the model's generalization. In addition to this, a lot of discussion has been done about the feasibility of the results produced from these studies. People argue that financial data follows a random walk and historic patterns and data can not be used to predict future prices. This *in theory* is supported by the economic principle of the Efficient Market Hypothesis (EMH) which states that market prices represent all current and past information already and hence there is no way to take advantage of patterns or mispricing to earn extra alpha [2].

To tackle the above problems, this paper aims to focus on a machine learning paradigm known as *Ensemble Learning*, use a set of various technical financial indicators as a feature set, and focus on emerging markets that tend to be less efficient.

Ensemble Learning refers to combining multiple learning algorithms to obtain a better predictive performance than could have been obtained from any of the constituent learning models alone [3]. Some of the common types of ensemble learning includes : boosting, bagging, and stacking. In this paper, we will be mostly focussing on stacking or *stacked generalization.* Stacking involves building different models to make predictions and then stacking them together and training a meta-model to learn from the outputs of the stacked models to make the best prediction.

In accordance with EMH, we are going to focus on cross market data with an emphasis on emerging markets that tend to be less efficient [4] and hence have relatively more opportunities to capture the alpha.

2. Objectives

The main objective of this project to come up with an ensemble model that is able to forecast buy/sell/hold signals and develop a trading strategy that is able to beat a simple buy and hold strategy for the most popular indexes for the geographies that are being targeted by the model. The minimum requirement for the strategy is to not lose money in relation to the Time Value of Money (TVM). The interest rate used to calculate the TVM will be the saving deposits offered by popular banks across different geographies.

The scope of the project involves :

- Feature Selection and Engineering : Experiment with different technical indicators and with different sectors across different geographies.

- Experimenting with different model types (SVMs, Decision Trees, Neural Nets) and/or different hyperparameters to optimize the ensemble.

- Validate the results using relevant statistical methods.

- Backest the trading strategy and compare the results against a simple buy and hold trading strategy.

3. Methodology

3.1. Problem Framing

The first step for any machine learning project is problem framing. Problem framing for this project includes thinking about the project from a finance perspective; thinking about the desired output, existing data sources, easily obtainable features, and existing challenges that our solution can tackle.

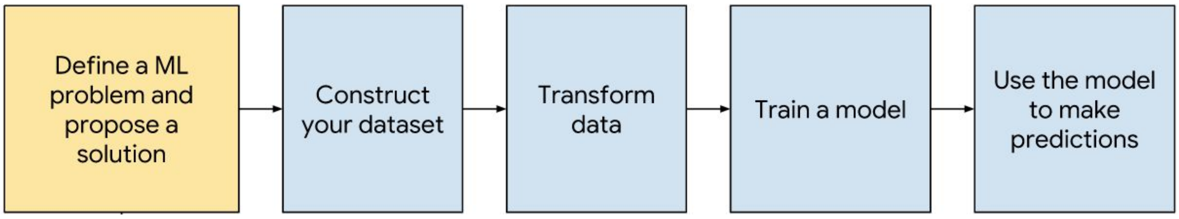
Define a ML problem and propose a solution
Construct your dataset
Transform data
Train a model
Use the model to make predictions

Fig. 1. A pipeline for a machine learning project [5]

3.2. Dataset

Finding good quality intra day tick data for markets is expensive and free sources tend to not be reliable. Hence, for our experiment we will mostly be using daily market data from Yahoo!Finance and Quandl. Both of them are reliable data sources and provide mostly structured data which makes data cleaning relatively simpler. The time period of the data will be from 2010-2019 (present day).

Data cleaning will involve making sure that missing data entries are filled by fill-forward or parametric sampling carefully, preventing future information from affecting present day values. The data will be indexed using stock symbols, geography, and time stamps. The data will be fed into models grouped by geography (cross-industry, same geography) and grouped by industry sector (same industry, cross-geography).

Data preprocessing will involve feature generation by calculating the various technical indicators using the price data. The parameters for certain parameterized technical indicators will be varied and tuned for different models and the best possible output. Also, the open/close prices fed into the model will be the adjusted open/close. An adjusted open/close takes into account various market phenomena for e.g. stock splits, dividends, etc. This gives a better sense of market conditions than just looking at an open/close price.

3.3. Ensemble Learning

To achieve the best possible results, different machine learning models with varying hyperparameters will be explored during the experiment. Some of the machine learning models that will be explored are as follows:

Support Vector Machines : Support Vector Machines or SVMs are supervised learning algorithms that analyze data used for classification or regression analysis.

Artificial Neural Networks : Artificial Neural Networks (ANN) are a collection of connected units or nodes called artificial neurons, which are based on the principle of a neuron in a biological brain.

Decision Trees : A decision tree (DT) is a decision support tool that graphs decisions and their possible outcomes based on different resource measures and utility.They are used to derive an items target value based on observations.[6]

3.4. Performance Evaluation

Various statistical methods (F-scores, Confusion Matrix) will be used to evaluate the performance of the model. Apart from them, as train-test split and k-fold cross validation do not perform well in the case of sequential time series data [7], walk-forward validation will be adopted.

The model will be backtested against historical data. Depending on time constraints either a backtest environment will be developed and set up locally or an online environment like Quantopian might be leveraged.

It will be considered a success if it is able to beat the market index returns of the geography targeted by the strategy.

4. Schedule and Milestones

- September 2019 :
    - Literature review about existing work regarding application of machine learning in finance
    - Research about various financial technical indicators
    - Explore data sources
    - **Phase 1 Deliverable on 29th September 2019**

- October - December 2019 :
    - Complete cleaning and preprocessing of data
    - Implement different machine learning models across different markets and sectors
    - Tune the models depending on the results
    - Further literature review

- January - February 2020 :
    - **First Presentation**
    - Develop a backtest environment
    - Tune the model depending on the backtest
    - Detailed interim report
    - **Phase 2 Deliverable on 2nd February 2020**

- March - April 2020 :
    - Finalize model and implementation
    - Draft final report
    - Final Presentation
    - **Phase 3 Deliverable on 19th April 2020**
    - **Final Presentation between 20-24th April**

- **Project Exhibition on 5th May 2020**

5. References

[1] Lukas Ryll, Sebastian Seidens *Evaluating the Performance of Machine Learning Algorithms in Financial Market Forecasting: A Comprehensive Survey*

[2] Investopedia : Efficient Market Hypothesis *https://www.investopedia.com/terms/e/efficientmarkethypothesis.asp*

[3] Wikipedia : Ensemble Learning *https://en.wikipedia.org/wiki/Ensemble_learning*

[4] Hesham I. Almujamed, Suzanne G. M. Fifield, David M. Power *An Investigation of the Weak Form of the Efficient Markets Hypothesis for the Kuwait Stock Exchange*

[5] Google : Machine Learning - Problem Framing *https://developers.google.com/machine-learning/problem-framing/*

[6] Wikipedia : Decision Trees *https://en.wikipedia.org/wiki/Decision_tree_learning*

[7] *https://alphascientist.com/walk_forward_model_building.html*