Capstone Project Proposal

Martin Piják

Bratislava, Slovakia

Abstract

This is proposal for using Machine Learning methods in futures trading.

Keywords: Machine Learning, Trading, Udacity, Nanodegree

1. Domain

In this project we will investigate futures trading on Chicago Mercantile Exchange (CME) markets with Machine Learning methods. The goal is to find the trading strategy mostly based on the price, Commitment of Traders report (COT) and seasonality pattern. We will compare this trading strategy to commonly used investing approaches as returns of Dow Jones Industrial Average or Bonds.

Example of Machine Learning used in futures algorithmic trading.

- Algorithmic Trading of Futures via Machine Learning
  - In this article you can find details about feature engineering, ML-algorithm selection and training
  - Input vector is 2 years of price and volume data.
- A Machine Learning framework for Algorithmic trading on Energy markets
  - This article deals with general pipeline setup for trading

2. Datasets and Inputs

I will use free data from Quandl.

Data range will be from 1st of January 2000. Until current trading day. On average trading year has 252 trading days. This gives us about about 4500 ($252 * 18 = 4536$) data points (these are composed of multiple values OHLC, COT, date, volume...) over past approximately 18 years.

I will retrieve data through Quandl API and then save resulting dataframe as CSV file. This way it will be possible to run project without the need to call Quandl APIs.

CME traded commodities below will be used.

- Gold
- Corn
- Coffee

I will use COT data joined with trading data (OHLC and volume) daily data.

I might look into other free data sources depending on the results I will get.

3. Data Analysis Tools

I plan to use following libraries/frameworks.

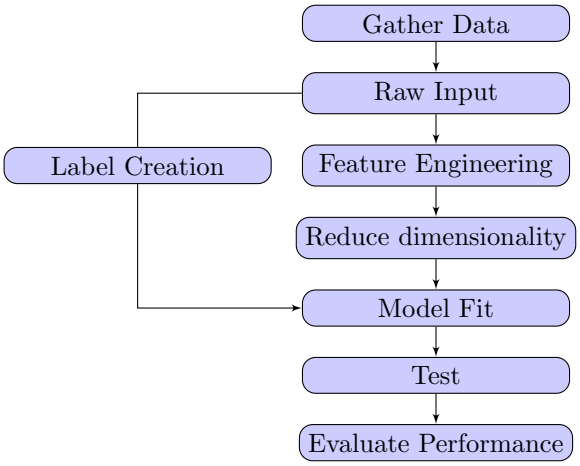
- Scikit
- Keras
- TensorFlow

4. Solution Statement

Main goal is to create trading strategy. I will follow classification approach.

4.1. Project Design

Following schema is showing high level project pipeline.



I will look into commodities that's why I want to capture seasonality. I plan to use daily data, because of that my trading plan is to hold the position (either long/short) for 1–3 days or until stop loss was hit.

I will investigate different features that have impact on trading.

- feature engineering
  - PCA (dimensionality of input vector is very high)
  - COT
  - price
  - volume
  - date (modify to find seasonality)
- training
  - Random Forest
  - Neural Network
  - XGBoost
  - LightGBM
  - ...

How will be stop-loss selected?

- fixed
  - try different fixed stop-loss values
- floating
  - based on the volatility of previous days

4.2. Raw Input

Raw input will have approximately following shape:

$$\text{2 years data}$$

$$2 * 252 \text{ trading days} = 504 \text{ trading days}$$

$$504 \text{ trading days} * 7 \text{ (number of features in a day)} = 3528 \text{ features}$$

4.3. Feature Engineering

Reduce dimensionality:

- PCA
- accumulation
  - trading day of month
  - average
  - try different min/max of last trading day in relationship to 2years of data
  - ...

4.4. Label Generation

Labels will be generated as,

$$fee = 1.5 \tag{1}$$

$$t_{treshold} = 30 \tag{2}$$

$$v_{volatility} = (\text{close} - \text{open}) * \delta \text{ where } \delta \approx 0.95 \tag{3}$$

$$labels \begin{cases} |v_{volatility}| > t_{treshold} + fee \implies \begin{cases} v_{volatility} > 0 \implies 1 \text{ (long)} \\ v_{volatility} < 0 \implies -1 \text{ (short)} \end{cases} \\ |v_{volatility}| < t_{treshold} + fee \implies 0 \text{ (no trade)} \end{cases} \tag{4}$$

δ constant is used for simulating slippage. 1.5 that we are subtracting is a trading fee. We can see in the label function if volatility is too close to 0 it will be rounded to 0 — meaning don't trade.

Output of model will be classification of input into three categories trade (long/short position) or do not trade.

If we will look into longer trading than one day. Labels can be constructed in the similar way as above. But we will have to consider stop-loss value. We might have to end the trade sooner than we wanted.

5. Benchmark Model

I will use three benchmarks for my model

- fixed percentage (bank/bond deposit comparison)
- Dow Jones Industrial Average performance
- mean reversal trading on given commodity markets

6. Evaluation Metrics

For simulation purpose I will calculate account gains with account size of $10,000$\$.

Trained models will be evaluated compared to Benchmark Models. Based on the one of trading signals $(-1, 0, 1)$ (output of huge input vector) gain or loss will be calculated — taking into account stop-loss. I plan to use 8 years for training and 2 years of data for model evaluation. I plan to trade only one contract regardless of the account size over time.