Machine Learning in Finance

Dragana Radojičić
Thorsten Rheinländer
Simeon Kredatus

TU Wien, Vienna University of Technology

October 27, 2018

Outline

The world of automation

- article from Washington Post: "The robots-vs.-robots trading that has hijacked the stock market", roughly 50% of all trading volume is executed by the robots.
- Stock markets are nowadays producing vast portions of data.
- The financial markets hold memory properties.

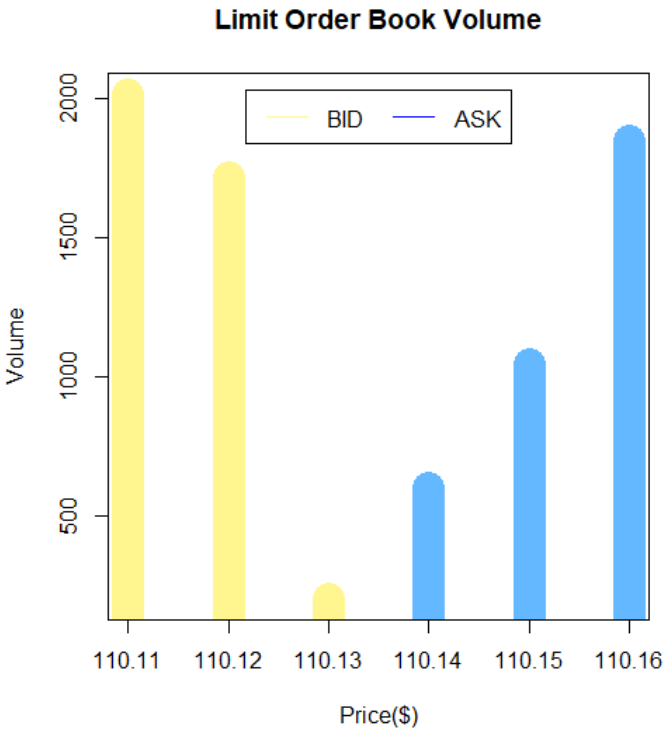
GOAL

The aims of our research is to analyze informativeness of the limit order book on future volatility and liquidity, in order to obtain further profit in the high-frequency trading.

Introduction: Limit Order Book (LOB)

List of all the waiting buy and sell orders.

- The LOB records all unexecuted limit orders.
- For a given price, orders are arranged in a FIFO stack.
- Tick is a minimal distance between two price levels (points in a discrete price grid).
- The spread is difference between the best ask and the best bid price.



The qualitative data analysis

NASDAQ (second largest exchange in the world)

- Our research is based on high-quality online limit order book data tool LOBSTER.
- LOBSTER has information for the entire NASDAQ stock exchange from the 27th of June 2007 up to the two days ago from the current day.
- 'orderbook' file - keep track of evolution of the limit order book
- 'message' file - contains information of the kind of event which update the limit order book (i.e. Time, Type, Order ID, Size of the order, Price, Direction of a trade)

GOAL

Goal is to develop a foundation which allows to easily match similar points together via unsupervised learning as well as to classify elements into groups via supervised learning (more precisely classification).

The data analysis

Label data

- The market data at a given time point t can be formally defined as a vector x_t, which will consist of market data informations and various technical analysis markers
- The main idea of our research is to express trader as a function with an input vector x_t such that output is one of the values from the set $\{S = idle, sell, buy\}$.

The classification, regression predictions and the latest research in the field of Artificial Intelligence shall be applied in order to successfully classify a time series of market data.

IDLE, BUY OR SELL?

Motivation of idea

- consider the real data history of Apple stock and to look at the conditional probability that Apple stock increase by at least 0.6%, condition on positive quarterly sales announcement.

$$p(\text{stockGoesUp} = 1 | \text{positive quarterly sales}) = 0.4.$$

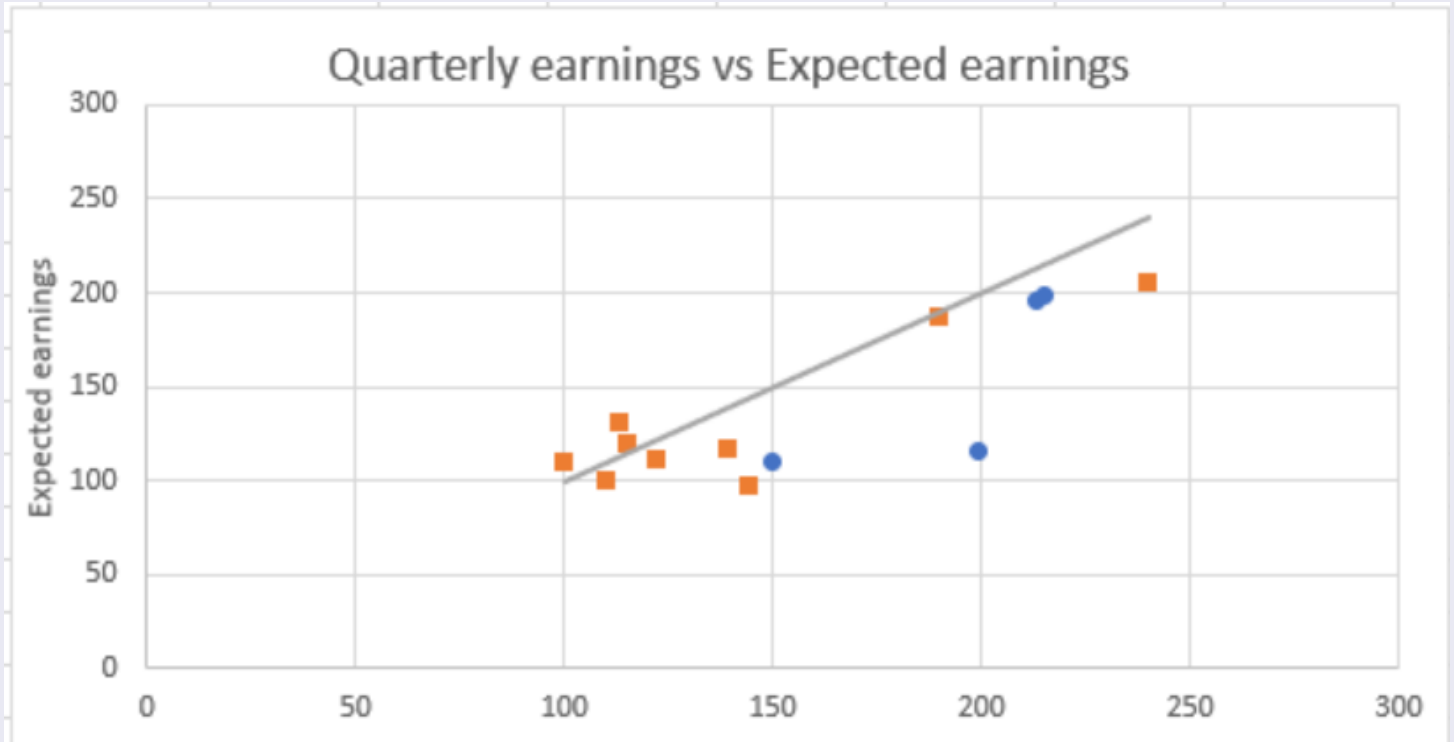
A sample of quarterly earnings plotted against the expected earnings.

Trading strategy which only indicates suitable time for opening a long position in terms of a Boolean function.

$$s(\text{qE}, eE) = \begin{cases} \text{True} & \text{if } qE \geq eE \\ \text{False} & \text{if otherwise} \end{cases}$$

A sample of quarterly earnings plotted against the expected earnings.

A sample of quarterly earnings plotted against the expected earnings.



The processing engine

The processing engine

- carrying out all of the data transformation
- enables researcher to prepare data on demand
- The pipeline itself consists of four major parts
  1. Raw aggregation
  2. Technical indicators
  3. Data lapsing
  4. Data labelling

Raw aggregation

Raw aggregation

During this stage a raw of a daily limit order book data consisting of the vectors is in the form

$$\begin{aligned} x_t = (&\text{bidLevel1}, \text{bidVolume1}, \text{askLevel1}, \text{askVolume1}, \\ &\text{bidLevel2}, \text{bidVolume2}, \text{askLevel2}, \text{askVolume2}, ..., \\ &\text{bidLeveln}, \text{bidVolumen}, \text{askLeveln}, \text{askVolumen}, \\ &\text{Time}, \text{EventType}, \text{OrderId}, \text{Size}, \text{Price}, \text{Direction}) \end{aligned} \tag{1}$$

Dataset

$$\mathcal{D} = \{x_t | 0 \leq t \leq \text{ amount of events per day}\}$$

Raw aggregation

aggregation function $a(\mathcal{T}_s^e)$

function which incomes $\mathcal{T}_s^e = \{x | x \in \mathcal{D} \ \wedge \ x_{time} \geq s \ \wedge \ x_{time} \leq e\}$ and outputs single vector y_s^e

complete partitioning over a single trading day

$\mathcal{P}_{interval} = \{\mathcal{T}_{j \cdot interval}^{(j+1) \cdot interval} | 0 \leq j < \frac{trading\ day\ duration}{interval}\}$.

- the set of aggregation functions $\mathcal{A}$, this set contains functions such as: extract minimum / maximum / first / last value of the vectors which depict the market order execution for each partition, etc.
- $\mathcal{I}$, as the set of all the intervals we want to get the aggregations from, the outcome of the stage is a set of datasets $\mathcal{Q}_a = \{\mathcal{D}_{interval} | interval \in \mathcal{I}\}$ where $\mathcal{D}_{interval} = \{(a_1(\mathcal{T}_k), \ldots, a_n(\mathcal{T}_k))_k | \mathcal{T}_k \in \mathcal{P}_{interval}\}$, $a_i \in \mathcal{A}$.

Technical indicators

- we need to recompute also further features providing insights about the market behavior.
- We use free open-source library called "TTR" (Technical Trading Rules) which provides the algorithm implementation for all standardly known indicators.
- define new partitioning $\mathcal{W}_{interval} = \{\{t_k | t_k \in \mathcal{D}_{interval} \wedge k \leq i\}_i | i \leq |\mathcal{D}_{interval}|\}$.
- define the new set of functions $\mathcal{M}$ consisting of functions $m_j(w_i)$ where $w_i \in \mathcal{W}_{interval}$ and output the technical indicator.
- the new set for each *interval*, $\mathcal{L}_{interval} = \{(m_1(w_i), \ldots, m_n(w_i))_i | w_i \in \mathcal{W}_{interval} \wedge n = |\mathcal{M}|\}$.
- At the end of the stage we define the feature enhanced set as $\mathcal{F}_{interval} = [\mathcal{L}_{interval}, \mathcal{D}_{interval}]$

Data lapsing

- during this stage we prepare a dataset which connects each interval with the related larger scaled interval.
- if $i, j \in \mathcal{I}$, and $i < j$, then the vector of features of the $\mathcal{F}_i$ will be joint together with the most recent interval j, which had closed prior to the start of interval i.
- we define the set $\mathcal{L} = [\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n]$, where $n = |\mathcal{I}|$.
- So the stage outputs the set of vectors, which are already enhanced by the feature extraction part and we can freely proceed to the labeling procedure.

Data labeling

- to be able to run any classification algorithm our training set needs to be labeled with respect to desired output.
- our labels are trades which we would like our algorithms to predict with respect to certain criteria.
- The most important criterion during the trading is to manage the risk and reward.
- we label all of the data points upon the fact whether we can reach certain profit with only exposing ourselves to certain risk until the end of each trading day.

Future work

- The main goal is to provide a common access for multiple users. Therefore the endpoint provides options for carrying out as much filtration and aggregation on the database level as possible and only get the data of interest back to the distributed engine.
- Study other interesting quantities.
- Model ORDER CANCELLATION.

Thank you for your attention!!!

Any questions?