BIG DATA ANALYSIS OF FUTURES: PENALIZED REGRESSION SPLINES FOR TRADE VOLUME PREDICTION AND PRICE VOLATILITY VS TRADE VOLUME RELATIONSHIP

by

Aniver Oluwatobi Bosede

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science in Engineering

Baltimore, Maryland

Abstract

Acknowledgments

Contents

List of Tables

List of Figures

Introduction

Technology has advanced so far that in our society today we are constantly collecting data [eco, 2010]. This has created the issue of how to feasibly analyze such overwhelming amounts of data [eco, 2010]. We can then consider how to efficiently store it or how to best carry out statistical computations [Zaharia et al., 2016]. The aforementioned together form the relatively new and evolving field of big data. The research here focuses on the analysis of hundreds of millions of rows of futures trading data with the aid of Apache Spark (Spark).

Spark is a fault-tolerant and general-purpose cluster computing system providing APIs in Java, Scala, Python, and R [Meng et al., 2016]. Spark was chosen for the analysis over the similarly popular Hadoop MapReduce (MapReduce) because of Spark's performance advantages as well as its greater computational capabilities [Zaharia et al., 2010]. Not only does Spark cache data resulting in persisted in-memory manipulations [Zaharia et al., 2012], it also includes Structured Query Language (SQL) and MLlib, a machine learning library [Meng et al., 2016]. Conversely, MapReduce only does manipulations via disk reads and thus does not allow for data sharing [Zaharia et al., 2016]. Furthermore for a typical pipline, external systems would have to be combined with MapReduce to provide querying and machine learning functionality [Zaharia et al., 2016]. Out of convenience the local file system was used for storage as opposed to a database (e.g. Cassandra) or HDFS (Hadoop Distributed File System). However, for industry data analysis, it would be worthwhile to invest the time needed to set up a more robust data storage system.

The futures trading data come from the Chicago Mercantile Exchange (CME) and were collected from May 2016 to November 2016. The futures were comprised of 22 financial products spanning six markets - foreign exchange, metal, energy, index, bond, and agriculture. First, this work uses spline regression to predict the volume of trading for any given day. Volume is taken to mean the number of total trades.

A spline regression was chosen due to lack of knowledge regarding the likely non-linear function underlying the true model. Predicting trade volume is of interest because many trading algorithms depend on volume [Satish et al., 2014]. Additionally, accurate volume predictions over a given interval allows traders to be more effective [Satish et al., 2014]. In general, volume prediction increases trading strategy capacity, controls trading risk, and manages slippage [Satish et al., 2014].

Second, the relationship between price volatility and trade volume is explored using standard deviation as a measure of volatility. In particular, volatility vs daily volume and volatility vs hourly volume was plotted to see whether or not the correlation remained the same with the passing of days and hours. It should be noted that there is evidence suggesting that comparisons between volatility and total volume do not extract all information [Bessembinder and Seguin, 1993]. This volatility-volume relationship is of importance due to the notion that hedgers are motivated to trade futures to stabilize their future income flows or costs, wherein the volume of their trading is based on their expectation of price variability [Foster, 1995]. Likewise, speculators are motivated to trade futures based on expectations of price variability [Foster, 1995]. Due to the fact that new information on the market causes agents such as hedgers and speculators to trade until prices reach a revised equilibrium which then changes price and trading volume, we expect a positive correlation between volatility and volume [Foster, 1995]. Indeed past research indicates that there is a positive relationship between volume and price volatility [Foster, 1995]. This sort of exploration provides information on the efficiency of futures markets which regulators can then use to decide upon market restrictions [Foster, 1995].

Methods

Results

Figure 1: Dummy figure 1

Figure 2: Dummy figure 2

Table 1: Dummy table 1

Discussion

Appendices

Bibliography

The data deluge, Feb 2010. URL `http://www.economist.com/node/15579717`.

Hendrik Bessembinder and Paul J Seguin. Price volatility, trading volume, and market depth: Evidence from futures markets. *Journal of financial and Quantitative Analysis*, 28(01):21–39, 1993.

Andrew J Foster. Volume-volatility relationships for crude oil futures markets. *Journal of Futures Markets*, 15(8):929–951, 1995.

Xiangrui Meng, Joseph Bradley, B Yuvaz, Evan Sparks, Shivaram Venkataraman, Davies Liu, Jeremy Freeman, D Tsai, Manish Amde, Sean Owen, et al. Mllib: Machine learning in apache spark. *JMLR*, 17(34):1–7, 2016.

Venkatesh Satish, Abhay Saxena, and Max Palmer. Predicting intraday trading volume and volume percentages. *The Journal of Trading*, 9(3):15–25, 2014.

Matei Zaharia, Mosharaf Chowdhury, Michael J Franklin, Scott Shenker, and Ion Stoica. Spark: cluster computing with working sets. *HotCloud*, 10:10–10, 2010.

Matei Zaharia, Mosharaf Chowdhury, Tathagata Das, Ankur Dave, Justin Ma, Murphy McCauley, Michael J Franklin, Scott Shenker, and Ion Stoica. Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing. In *Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation*, pages 2–2. USENIX Association, 2012.

Matei Zaharia, Reynold S Xin, Patrick Wendell, Tathagata Das, Michael Armbrust, Ankur Dave, Xiangrui Meng, Josh Rosen, Shivaram Venkataraman, Michael J Franklin, et al. Apache spark: a unified engine for big data processing. *Communications of the ACM*, 59(11):56–65, 2016.

Curriculum Vitae