Bachelor Thesis Review

Faculty of Mathematics and Physics, Charles University

Thesis author	Martina Fusková
Thesis title	Anomaly detection for stock market trading data
Year submitted	2020
Study program	Computer Science
Study branch	General Computer Science

Review author	Filip Kliber	Reviewer
Department	Department of Distributed and Dependable Systems	

Overall

	good	OK	poor	insufficient
Assignment difficulty		X	X	
Assignment fulfilled			X	
Total size *. . . text and code, overall workload*			X	X

The goal of this thesis is to study various methods for detecting anomalies in stock market trading and implement these methods (models). The goals of the thesis have been fulfilled, however there are some drawbacks.

Thesis Text

	good	OK	poor	insufficient
Form *. . . language, typography, references*			X	
Structure *. . . context, goals, analysis, design, evaluation, level of detail*			X	
Problem analysis		X		
Developer documentation		X		
User documentation		X	X	

Textual part of the thesis is written in english, but used language style makes the thesis a bit harder to follow, mostly due to higher occurrence of grammatical errors or sentences with a missing word (e.g. *For us was the core functionality of this library the ndarray.*; or *In sklearn can also be found functions for pre-processing*). Sometimes author uses different wording for (most likely) the same concept, which makes the section confusing (e.g. in Section 1.5.7: *Standard neural networks*, *Deep neural networks* and *basic neural networks* are referring to the same thing here). Overall structure of the thesis seems reasonable.

In the first chapter, author provides theoretical background in machine learning, feature engineering, statistics and also presents various models used for detecting anomalies in data. However, the models are not divided or categorized in any way (even though, later in the thesis, they are divided into three categories).

Second chapter is dedicated to the analysis of the problem. After selecting python as a programming language for the implementation of models, this chapter mostly consist of discussion about various libraries used in machine learning and anomalies detection. In only handful of these cases, a solid argument is presented to support the choice. Furthermore, there is no introduction to the FIX protocol.

(pokračování na další straně)

Last part of second chapter describes what issues the author encountered while implementing the models within the application, but doesn't provide the solution to these problems in all cases. For example the author states that *For models that should make predictions based on several days and the time window was small, their development and training took even several days.* and the paragraph ends there and lacks any clue on how was this issue addressed.

The following chapter is supposed to present architecture of the solution, but author very quickly jumps to technical and implementation details without giving a bigger picture. Later in this chapter, in section 3.2.6, author states that *Logically we can split the models into two types – forecasting models and density models*, but follow up sections also describe *Seasonal models.* That could be viewed as a simple typo, if it had been explained beforehand.

Fourth chapter explains the process of experiments and the results. The explanation is quite vague, as I didn't understand the results. Author used the application on real, existent broker data from a 2018 and found some anomalies in it, but also state that *anomalies were injected into the last log by an algorithm created by an expert.* What anomalies were found then?

Author grades models based on two performance metrics — TPR (true positive rate) and FPR (false positive rate), which seems like a reasonable idea, but this hides a fact that combining a usage of multiple models detecting mostly distinct anomalies would have a higher score. Author doesn't provide any details on found anomalies, such as whether some are found by all models, or if there are rare ones found only by few of them.

At first glance, the textual part seems to be sufficiently long (72 pages), but if we omit formal requisites, theoretical background and images of graphs, the thesis would be about 25 pages long, which seems a bit too short for a bachelor thesis.

Thesis Code

	good	OK	poor	insufficient
Design *... architecture, algorithms, data structures, used technologies*		X		
Implementation *... naming conventions, formatting, comments, testing*		X		
Stability		X		

The implementation part of the thesis consists of the implementation of two programs in Python language — one for extracting features from stock market log file in FIX format to be used in machine learning algorithms; and other for detecting anomalies by using various models described in the thesis, with the help of machine learning.

The source code is readable and easy to follow, but I am concerned about the size, which is about 58KB for the feature engineering part and 42KB for anomaly detection part. The implementation of anomaly detection however mostly just calls the appropriate models from libraries and frameworks for machine learning (with some exceptions, for example the implementation of Kalman filter). The source code is well documented.

The user manual doesn't give any insight on how to install the application. The package contains `requirements.txt` file, but it is not in the correct format for python's `pip` installing machinery, thus I had to adjust this file to perform the installation of required dependencies. A command used to run the application is provided as a screenshot of terminal instead of a text that could be copied. What is also a bit inconvenient is that both of the programs expect the input (1–2 file names) to be passed on standard input, instead of on the command line.

Overall grade Good
Award level thesis No

Date Signature