BSc (Hons) Information Systems with Business Management

Final Project Report

Clustering Based Outlier Detection for Financial Data Analysis

By

Ms. SUBASHINI DILSHIKA NISSANKA

2013529

m.nissanka@my.westminster.ac.uk

Supervised By

Mr. ALROY MASCRENGHE

This report is submitted in partial fulfillment of the requirements for the BSc (Hons) Information Systems with Business Management degree at the University of Westminster

Abstract

As the world grows in complexity, we are overwhelmed with data. The more the volume of data increases, the more the proportion of people understanding on data decreases. People understanding on data is a big concern these days as data adds more value in organizational context by producing new insights, inspiring innovations, allows more distinct decisions at correct time to gain above average competitive advantages. This big concern is not a big deal anymore with statistical data analysis. There are plenty of techniques, tools, algorithms which analyze data, to make people see, understand and share analyzed data for better performances. Therefore, having realized that people should study statistical data analysis and data visualization to explore transparency on data to unveil hidden stories and to create knowledge.

To this end, the aim of this project is to study, whether an unsupervised data analyzing (machine learning) approach would be feasible to detect suspicious data or outliers in financial data. Outliers are patterns in data, that do not conform a well-defined notion of normal behavior. In this project, I focus on data clustering technique for outlier detection. Clustering is an unsupervised learning approach which is commonly used for statistical data analysis. Clustering algorithms are powerful and group massive datasets into sub-groups which share similar characteristics (clusters) quickly.

The project proposed a software tool called DeepScan to detect outliers in financial data. DeepScan allows users to analyze financial data to detect abnormal occurrences to avoid financial losses and poor organizational conditions. A cluster visualization and a statistical report are the outcomes of DeepSan and it allows user to decide whether the detected data instances are belonging to an outlier in their business context or belonging to a normal category which has novel characteristics. This project also validates the proposed solution-DeepScan achieves impressive results.