

Vaughan.Mark.Final Report

by Mark Vaughan

Submission date: 13-May-2022 02:49PM (UTC+0100)

Submission ID: 179934373

File name: 19487_Mark_Vaughan_Vaughan.Mark.Final_Report_398307_1879735091.docx (1.47M)

Word count: 3300

Character count: 17363

Document Title

Assignment

	Mark L. E. Vaughan Undergraduate, University of Wales Trinity St David Swansea Wales 1909160@student.uwtsd.ac.uk	

Contents

1. Introduction.....
2. What is Data Mining.....
3. Data Set Description
4. Processing Description.....
 - A. Pre-Processing
 - B. Post Processing
5. Sample Runs
6. Business Recommendations.....
7. References & Bibliography
8. Appendices

Abstract

1. Introduction

2. What is Data Mining

Observation of figure one will provide a root definition of data mining. While this holds true in principle, the reality is that the database is for the

Data mining :

Overview

Examples

Definitions

Definitions from Oxford Languages · Learn more



data-mine

/ˈdeɪtəmɪn/

verb

gerund or present participle: **datamining**

analyse (a large database) in order to generate new information.
"users data-mined the new, free content earlier this week"

Figure 1

ion extraction as part of a data warehouse or lake. Since 1974 relational databases have been the staple of information storage in all aspects of

human activities and as with a lot of technological systems the principles of chaos have been at work altering the requirements and need of systems. This change is answered by the development of new methods and techniques in information storage. One of these methods typifies the need for datamining. This is graph technology. Data is useless without a reference or meaning to transform it into information. This reference or relation is what powers referential data systems. The appearance of Graph technology provides meaning in a far greater way with the ability to link facts (a table of attributes) together in a wider context. Current relational databased storage by its nature often hides these connections and relies on data mining tools such as Weka to use machine learning techniques like clustering and varying forms of classification to predict and discover these hidden linkages and generate new information for the users of the system. The next big challenge will be to apply these tools to the multidimensional constructs of Graph based data webs, information constellations and n-dimensional globular clusters.

3. Data Set Description

The dataset used is the Online Retails dataset from the UCI archive held at the University of California Irvine. The dataset is a retail record of the line-item entries which form the bill of materials for the invoices raised by the merchant. The bill consists of the descriptions for the stock item, its stock code, the unit price and the quantity delivered. Each record also contains the invoice number, the customer code, the invoice data, and the country of origin for the order. In its raw form the dataset is flawed. This is evidenced by the entries that were made in contravention of the rules of accounting and the principals of double entry bookkeeping. Errors include using negative quantity numbers, manual overwriting of stock descriptions, missing customer account ids and incorrect use of discount entries. The dataset also shows evidence of

cancellation invoices which are not an authorised business practice. These entries should have been entered into the ledger as credit notes and not as invoices to keep the credit and debit balances correct.

4. Processing Description

To render the dataset suitable for processing, the data needs to be cleared of errors and obvious out of range values often referred to as outliers. Prior to pre-processing, several small errors of omission were corrected, these were details like inserting customer ids into records where they were missing from records where in identity were known.

A. Pre-Processing

With the minor errors of omission removed, pre-processing proper commenced. At this stage a choice of methods was considered. The first choice was to remove all records with incorrect format or missing values. The remaining records being the baseline data set. The alternative method was to treat the errors as a form of data and mark the affected records accordingly. It should be noted that the classification was done manually due to the degree of variation of the text errors. This was particularly noticeable with description overwrites. The zero currency and negative quantity entries were easily marked with a add attribute filter with split values of zero. A similar filter was used to identify all currency values greater than 10,000 and less than -10,000. The arbitrary value of ten thousand being chosen to accommodate any wholesale purchases as there is no wholesale identifying code in the customer ID coding.

For the assignment the latter method was chosen with nine additional binary attributes added to indicate the nature of the error. See the baseline data file for details or view the header as shown in appendix 1. Stage two was to alter ambiguous field entries to dedicated text values and had any numeric vales be found missing, the record would have been deleted.

B. Post Processing

With Pre-processing concluded, the next stage is to prepare the dataset for practical use. This is achieved by setting a baseline classification reliability value, creating a validation subset that will be used to determine the effectiveness of the classification process by using untested data and to create versions of the cleaned dataset which would convert selected nominal attributes to numeric equivalent s and vice-versa.

Using the Weka workbench tool. The first step in postprocessing was to generate the validation set from the cleaned data to permit the testing of results from untrained data. As the data would never be used for classifier training, results from classification processes anomalies from the training data will not affect the results. In using this data, deviations from the validation and testing data can quantify the accuracy of test predictions made.

The baseline reliability value will use the ZeroR classifier which uses the general statistical probability without bias from the incumbent data. For more detail I would advise viewing class 2 Lesson 4 of Professor Ian Witten's Tutorial course[1] on you tube. Using this classifier will set the lowest feasible value for accuracy for the dataset. If tested results fall below this level, an indication of serious classification settings are indicated, and a revision of the determinate hypothesis should be made. The use of the baseline would also give a good indication of the effectiveness of the classifiers used by evaluating the degree of variances found.

Lastly, Variant and sample data sets need to be created and processing filters created. This was achieved using the workbench tool built into Weka. The filters can be found within the built-in function within the program. For further detail lesson 5 in class one and lesson 3 in class 3 of the beginner's tutorial[1] will be illuminating.

5. Sample Runs

With the refined dataset ready, evaluation set on standby, and a baseline reliability percentage found

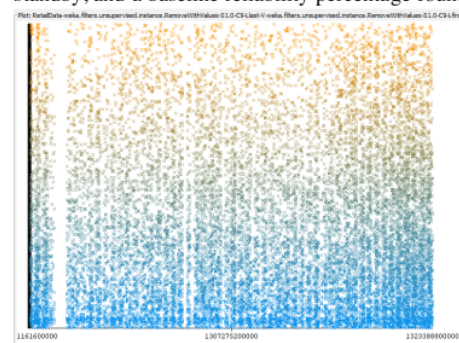


Figure 2 Seasonal Distribution

initial investigation commenced. The first undertaking was to load the baseline dataset and checked the visualisation matrix for any noticeable patterns. On observing the matrix, a distinct curve was detected on the invoice No/ date of invoice chart. This was consistent with the increasing invoice number with time however, it was noticed

that the plot was obscured by the number of UK datapoints. For greater clarity, two new datasets were created using the Remove Instance by Value filter to separate the UK results from the baseline dataset. This having done, the two datasets were opened in the Weka Explorer and viewer side by side in the visualisation matrix. Two further observations were made in the invoice Date/CustomerID, the first was a regular spaced series of gaps along the date axis. On checking the dates bordering the gaps it was determined that the gaps were the dates for Sundays thus indicating a six-day working week is operated by the company. The second observation was distinct concentration of datapoints along the customer ID axis of the chart. The presence of these lines was interpreted as a series of constant and regular orders placed by a particular customer. It was also noted that there was a large gap in the data around the end of March. As the gap is present in both the UK and foreign datasets, the cause must be related to company operation, possibly computer system failure or a scheduled maintenance shutdown. Using the visualiser for invoice Date / CustomerID the concentration at the start and end of the series indicates increased ordering in line with the Christmas season. Also notable are the slightly wider weekend breaks around the end of May, June and August which co-inside with UK bank holiday dates.

It was at this point that analysis test runs were to be performed. However, a hardware fault on the testing system caused Weka to freeze during the calculation phase of the analysis algorithms. The technical term is Zombie Process failure. Visual evidence for this can be found in appendix 3. Some preliminary data was gleaned before the hardware faults occurred and these can be found in Appendix 2.

The original first run was to use a subset of the cleaned database created using the random sample filter on the workbench to determine profitable, ascending, falling and loss-making stock lines. The string attributes were then converted into nominal attributes and all instances with quantities of five items were removed to eliminate wholesale orders. For this the remove by value instance filter was used. The forecast analysis was to be generated by using an AddClassification filter using a logistic regression classifier as the tool to determine projected values for the stock lines. This would determine whether the line would have increased, stable or decreasing order quantities for the near future. The results of the filter would appear as additional attributes in the dataset which could then

be filtered for the desired values. Further uses would be to invert the selection processing to identify wholesaler stock lines and their current sales characteristics. This could be used to identify potentially yet undiscovered stock lines using the multinomial Bayes Text Classification to cluster stock lines with similar text descriptions. This later classifier would require conversion of text nominals to string format.

The second set of runs were to be descriptive in nature using clustering techniques to analyse the numeric attributes in the data sets to correlate quantity/unit value correlations against nominals such as country and CustomerID to derive buying profiles and sales trend for customers and exports. The analysis when used against invoice dates could also be used to quantify the seasonal trends within a market and indicate which trend to concentrate on. Weka provides a decent range of filters for this kind of descriptive analysis. For these runs the main choice would be to use the Simple K-Means as it is basic and uncomplicated to implement. Other Cluster algorithms considered are E.M. and DBSCAN these are widely used and are well tried however, with only ten days to learn the system, knowledge of the finer adjustments to these algorithms has yet to be learned. Unfortunately, with the hardware failure, this will have to wait.

* In this paragraph I was going to talk about the degree of pre-processing and configuration of datasets for refining data models. But time and word space has run out, so I left this note...

6. Business Recommendations

With examination of the available data the following recommendations are presented to improve company operation and profitability.

1. Complete revision of accounting and invoicing processes.

As indicated earlier, the raw data indicates a lack of understanding of the accounting process. The use of cancellation invoices should be abandoned. Double entry bookkeeping works by tracking the movement of funds between accounts as a credit or a debit. In this scheme, invoices are statements of goods and services provided and are a demand for payment. These in accounting terms are classed as income which reduces the debt on the customers ledger. Operation of invoices are strictly credit based only. They cannot be cancelled or altered after posting. To do so would be a violation of the rules of accounting and illegal. Where corrections are needed, corresponding credit notes should be

raised and used as a payment entry in the ledger to offset the invoices charge on the customers ledger.

The practice of amending stock item descriptions is also to be avoided. Damaged stock should be handled by transferring the asset to a separate ledger using stock transfer documents and not on a public bill. The presentation of which would indicate a state of incompetence. The practice is also used for billing unrelated items using the stock code assigned to a particular asset. This would cause inaccuracies in the stock keeping process and is indicative of a poorly designed stock schema.

Lastly, the absence of customer ids and stock codes in the raw data hints at poor data input controls and poor data management rules. No invoice should be posted where there are missing field entries. If the log of invoice postings has missing entries, then there is a problem with handling of the data within the operating database.

2. Introduction of stock and customer coding schemas.

Investigation of the stock codes indicate that there is no apparent coding scheme built into the issue of stock codes. The use of such a coding scheme would make inventory control and analysis easier. From the datamining perspective, the use of codes would increase the flexibility of classification allowing for the creations of subcategories within a single attribute. Mining with this capability would increase the accuracy of algorithms and reduce the incidence of error within data models. Similarly, a coding scheme for customer ids would provide an easier separation of public and wholesale accounts. As buying patterns for each group are contradictory, separating these groups as part of pre-processing would be preferable and a definitive coding scheme would be beneficial when applying instance filters to datasets.

3. Creation of a preferred customer scheme.

The application of datamining techniques can be leverage for the creation of a preferred customer list. This list would be based on parameters such as ordering activity, periodic purchase totals and the range of lines ordered per purchase. By classifying purchase records using combinations of these three parameters, clients that trade intensively with the company can be identified. A recommendation would be the creation of additional discounts, lower cost per volume purchasing prices and extended credit facilities for clients on the mined list. This feature would increase brand loyalty, boost trading volumes, and build improved trading relationships.

Increased profitability could then be applied to general accounts as extra services such as gift wrapping and priority delivery.

Purchasing and inventory management recommendations would be to discontinue the bottom five percent of sales performance on a quarterly basis. Had the first run been able to run properly, the results would have provided a twenty-bin classification based on the line totals per instance. This would show (simply) the relative profitability of stock lines. Those items that fall into the lowest class of this list would be removed from the procurement list and those in the highest class would be prioritised for stock acquisition as extra stock holding for the quarter. Adjustments would then follow as ongoing performance reviews continue.

7. Reflection

A brief note about the experience. Having used the Weka software, its function is simple and well thought out. The documentation needs some work and a better laid out descriptive format would do much to improve the usability. The tutorial shows how to run filters, classifiers and clutterers with the training data but gives little by way of explaining the parameters and algorithm requirements. It took three days to realise that string attributes need to be converted to nominals before the bulk of the classifiers would become usable.

While unfamiliarity with the software caused the learning curve to be steeper than it should have been, the observation concerning data mining and warehousing is the lack of clarity concerning the terminology around the subjects. Put simply the subject is chock full of synonyms. Data warehousing has dimensions and measures, data mining has nominals and numerical. In function and attribution, these are identical. One is a set of text values, and the other is a single numeric value. Dimensions/nominals can be used as the basis for classification in a deterministic model with the numeric/measures being used to set scale and magnitude to the determinate class. Add to this confusion statistical terminology and you have an ironclad method of keeping the aspirin manufacturing industry in lifetime employment. It would be simpler to align these varying terms into a single scheme of description so that things like applying classification techniques to a hypothesis test model is more transparent and easier to understand.

This aside, the current state of mining technology advanced considering the use of flatbed data tables

as the primary data source not that there is anything wrong with this form of structure. Like anything concerning data manipulation and transformation, the quality of the result is only as good as the raw material used. Therefore, data prep is important and doubly so where warehouse-based data is used. Much of the learning experience was in learning to isolate the key attributes and which secondary attributes are needed to obtain the best accuracy with creating data models. Weka uses a filter system to allow manipulation of base data to adjust and alter the data structure and values to optimise the data set but what is not obvious is the inherent danger in over fitting training datasets to get the best results. From this standpoint the use of fixed data models is not an advantage. Training schema should be based more on an iterative Bayesian technique where mining models should be built on complete inclusive datasets in batch mode allowing the model to evolve as the training progresses. This would permit very large data sets to be processed on more modest hardware systems and would normalise the entire dataset eliminating errant value distortions which would normally skew the results.

As part of the emerging trends module, I investigated future trends in database and information system technology. With this knowledge, the effectiveness of data mining will explode when this technology is integrated into graph data technology. The fundamental basis of graph technology is the ability to create links and connections between discrete items of data. The primary role in data mining is to discover links, patterns, and connections. With the two technologies combined, the exploitation and capability in forecasting, resource management and knowledge discovery will have far reaching consequences as the ability to combine datasets from any field of endeavour and analyse them into a matrix of possible models is made easier and more interactive.

8. References & Bibliography

- [1] "Weka 3 - Data Mining with Open-Source Machine Learning Software in Java." <https://www.cs.waikato.ac.nz/ml/weka/courses.html> (accessed May 06, 2022).

"Data Mining Techniques and Applications", Du Hongbo, Cengage Press, ISBN:81-315-1955-4

9. Appendices

C. Appendix 1.

Header for the baseline data file.

```
% File: BaselineData.arff
% Author: M. Vaughan
% Created: 14 Apr 2022
%
% Arff data file for Data Mining Assignment. Augmented with additional binary numeric attributes for error analysis and filtering.
%
% Extended Binary Attribute Description
%
% M1 - Outlier Values
% M2 - Reg. Quantity Values
% M3 - High Discount Quantity
% M4 - Manual Entries
% M5 - Invalid Description
% M6 - Multiple Code Description
% M7 - Cancellation Invoice
% M8 - Zero Value Adjustments
% M9 - No Customer ID
%
% Relation RetailData-weak.filters.unsupervised.instance.RemoveWithValues-01.0-09-Last-9
%
% Attribute InvoiceNo string
% Attribute StockCode string
% Attribute Description string
% Attribute Quantity numeric
% Attribute InvoiceDate date dd/mm/yyyy
% Attribute UnitPrice numeric
% Attribute CustomerID string
% Attribute Country (Australia,Austria,Bahrain,Belgium,Brazil,Canada,"Channel Islands",Cyprus,"Czech Republic",Denmark,EIRE,"European Community",Finland,France,Germany,Greece,"Hong Kong
% Attribute M1 numeric
% Attribute M2 numeric
% Attribute M3 numeric
% Attribute M4 numeric
% Attribute M5 numeric
% Attribute M6 numeric
% Attribute M7 numeric
% Attribute M8 numeric
% Attribute M9 numeric
%
% Data
```

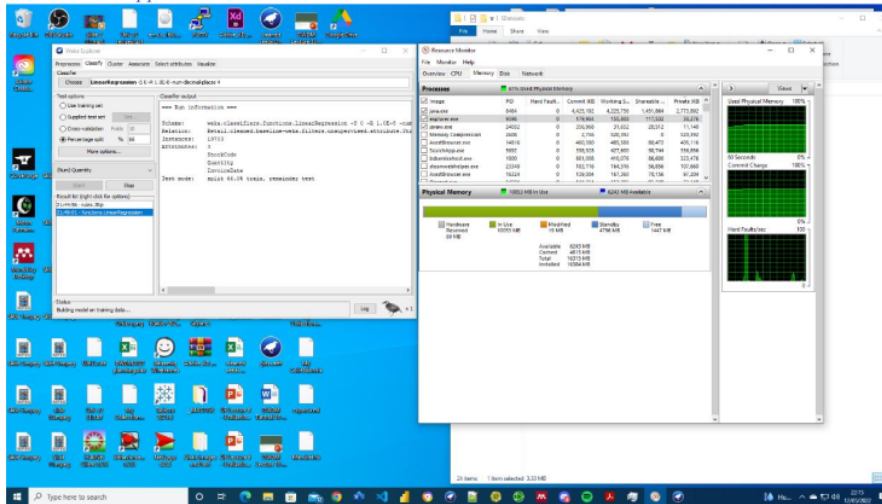
D. Appendix 2.

Assignment Datafiles can be Accessed online at:

<https://1drv.ms/u/s!AgT4R4ITrV6pg5tadeixELDuMp6wng?e=t33VQW>

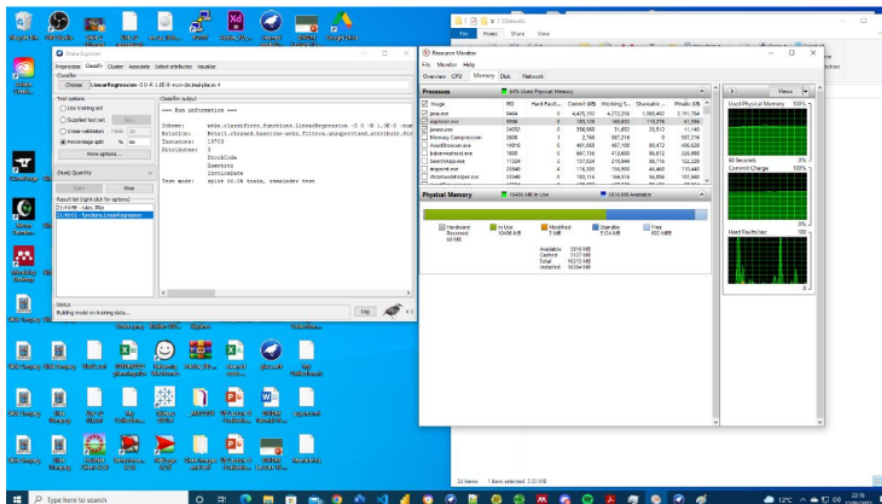
No password Required. Please note Access is Read Only.

E. Appendix 3.



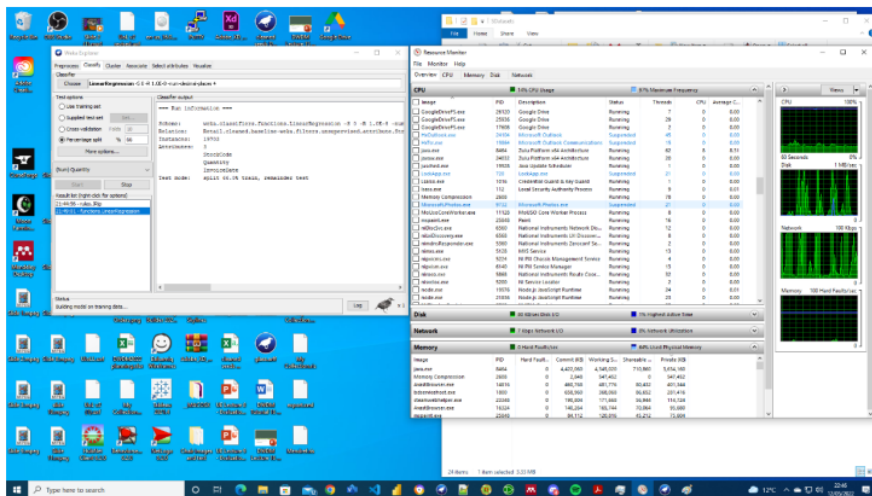
Screen shot 1.

First indications of memory controller malfunctions. Note the spikes in the Hard Fault display in the process manager. Data from the still active Java applets is not being saved to memory as the memory controller is not responding or the memory chips used have bad storage elements. Most likely the data sent would be missing or corrupted. Also note that the process manager shows standby memory is available indicating that this is a hardware fault and not a software glitch.



Screenshot 2.

Taken one minute later, the process manager is still showing spikes in mamory transfers at the hardware level. This shows that the fault is percistant and not transient. This will require the replacement of the DIMM modules but not in time for the assignment deadline.



Screenshot 3.

Note: the reduced number of Hard Fault instances in the memory panel and that the CPU load on the Java applets are minimal or at zero indicating that the applets are still in memory but have little or no CPU activity. They are dead ergo Zombie Processes.

FINAL GRADE

57 / 100

GENERAL COMMENTS

Instructor

Mark Vaughn	
1 engagement with literature comment	available
Only one reference is provided - this is cited within the body of work but is far too brief for work indicative of level 6	20
2 Knowledge and Understanding comment	available
Very good discussion that denotes an understanding of the tools and techniques demonstrated throughout (this should have been reinforced with background research)	10
3. Cognitive and Intellectual skills comment	available
Good discussion of results but little evidence of the results included within the report	25
4. Practical Application Skills comment	available
good discussion and justification of the use of weka and the stages to complete the experiment, however, little illustration of how this would be completed (system issues identified that may have impacted the ability to complete the investigation)	35
5. Transferable Skills for life and Professional practice comment	available
Nicely laid out columnar report	
Incorrect title	15
avoid informal writing and first-person narrative	
Total:	57

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8
