ZA review of commonly used data mining Tools

A compare and contrast of 3 leading data mining tools

Bhavneet Soni

Department of CSIS

Harrisburg University of Science and Technology

Harrisburg, USA

bsoni@my.harrisburgu.edu

*Abstract* — We reviewed three data mining tools namely Enterprise Miner Statistical Analysis System (SAS), Statistical Package for the Social Sciences (SPSS) by IBM and XL Miner. Upon comparing and contrasting these tools we find out that SAS provides a complete set of statistical and analytical tools, but SPSS is easy to learn and should be considered if starting to learn data mining. XL Miner provides a familiar interface as most people have used Microsoft Excel and provides a relatively cheaper alternative that can be easily implemented.

Keywords—data mining; SAS; SPSS; XL Miner;

# Introduction

For this assignment, we referred to a paper by Haughton et al [1], paper reviewed five of the most widely used and popular data mining tools. Data mining a term first coined in 1990’s is a process of extracting patterns and meaningful information from large data sets [2]. Data mining implements principles and fundamentals of statistics, artificial intelligence, computer science and other similar fields of study. Three of the main data modeling we will compare in this paper are SAS, SPSS and XL Miner. Reason for choosing to review these was mainly that SAS and SPSS are one of the most popular and complete statistical packages, whereas XL Miner is an additional extension on to MS Excel which almost everyone in familiar with and have used in their life time at one time or another. These tools were compared based on many factors such as cost, statistics and graphics capabilities, predictive models used and association analysis. Although the paper was written back in 2003, we have researched and put present day capabilities of these tools for comparison.

Author used different data sets to analyze different aspects for the tools. For the sake of descriptive modeling analysis Direct Marketing Education Foundation dataset 2 along with Census geo-demographic variables dataset6 acquired from [www.the-dma.org/dmef](http://www.the-dma.org/dmef) was used. Data consisted of 19,185 data points for a business with multiple divisions that were mailing out different catalogs to clients in the same customer database. Whereas for association analysis Bookbinders club case dataset from Direct Marketing Educational Foundation. Data set consisted of 1580 customers.

# Cost and pricing

While deciding for which tool to buy one of the most important criteria for any enterprise or an individual is the cost. We try to compare the market price of these tools and licensing options offered for the product. Although the author discussed the prices of various software and licensing for these in 2003, I think with the rate of change of technology and costs, those numbers are irrelevant and tried to find out what these costs today. As per recent estimates IBM SPSS modeler costs about $4,400 for a single user personal license to about $11,000 for a single user Premium license [3]. SAS single user license for SAS Analytics Pro cost about $9,380 for desktop version whereas cloud based versions cost from $6,100 to $9,180. XL Miner single user first year license cost from $11,000 to $250, depending up on the package, with renewal licenses costing up to $5,400. However, since XL Miner is an add on for Microsoft Excel, having an Excel license is a prerequisite for using XL Miner. While these are the numbers available on the web, these companies offer various prices based on the number of licenses and packages being bought. Each of the products are available for Enterprise and educational licenses but more and more companies are moving towards cloud based computing and cloud based services where one can use the tool without needing to buy the tool and hardware to host and run them.

# Descriptive statics and graphics

Descriptive statistics and graphics are the first and foremost measure of a data set. Descriptive statistics gives us a measure of the central tendency of a data set such as mean, median mode etc, it also gives us the measure of variability of the data by range, standard deviation and variance etc. [6] Whereas graphics capability gives a measure of how best the tools can represent the data in a graphical format such as scatter plots, histograms, bar graphs etc. These parameters are commonly used to gain a better insight of the data before any analysis or modeling could be initiated. SPSS provide an easy to use and find missing data points by way of “Quality Report”, it also offers rich techniques to treat the missing data points that can be easily run. SPSS and SAS can handle large data sets, XL Miner is somewhat limited in handling larger data sets, it can handle 2000 rows with 200 columns at a given time with no more than 6000 records, these numbers are from the time paper was written but it gives an idea that it’s not as good as other tools with large data sets.

SPSS provide excellent graphics and user interface that provides extensive documentation and help topic just by hovering over the various tool icons and buttons. SAS also provides for good help and documentation but XL Miner seems lacking in this section. Each of the functions and tools are represented by icons and buttons on the screen, these can also be accessed from the drop-down menu items. SPSS provide several types of graphs which are very easy to plot such as scatter plots simply called as plots, distribution plots, histograms, collection, multi plots and web plots. Web plots use line thickness to represent overlapping of categorical values. While SPSS does not have, the standard box plot it provides Web graph, as shown in Figure 1, it’s a unique tool that’s only being offered by SPSS. These tools provide option to convert and export the plot graphs in .jpg or .png file formats. SPSS stands out from the pack for its publication quality graphics. SPSS also differs from XL Miner in the way that it allows for evaluation feature that can specify data based on percentiles allowing for easily modifying the graphs.

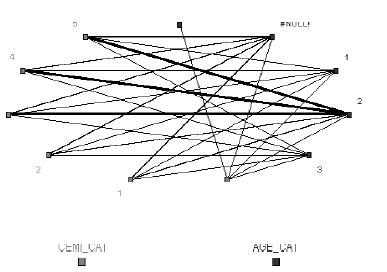


Figure 1. Web Graph in SPSS

SAS provides a wide range of tools for preliminary analysis, Enterprise Miner Network can work with raw data as well take in the transformed data produced by various data transformation modules such as Replacement and transform variables. SAS and XL Miner does not have multiplot graphs that are available in SPSS. SAS provides a high degree of complex analysis even before high end modeling is initiated. SAS provides for 3D rotating plot and interpolation of contour lines which are very helpful in visualizing data sets these are of best case scenario for implementing weather data analysis. While SAS offers, rich graphics package its user interface seems lacking. There is not much flexibility in terms of modifying and manipulating the graph characteristics. SAS also provide distribution, variation and curve fitting tools that helps in having a quick initial analysis of the data set. It can help predict trends and multi variable regression analysis

While SPSS can directly use CSV, SPSS and SAS files it needs other formats (Excel and dbf) to be transformed before it can use them, whereas XL Miner can easily use them and have inbuilt wizard support to transform other forms of data into usable formats. SAS also provide import wizards to transform data types into usable formats.

# Predictive models

Predictive modeling in layman terms mean forecasting, several variables and factors are considered to predict a future outcome based on the collected data. The method can implement simple linear equations to much more complex neural networks for predicting [7]. A model is trained by a small data set and then the model is validated and refined based on the availability of more data. SAS stands out from other tools in this section as it can perform all the steps required for setting up a predictive model and training it automatically.

SAS Enterprise miner takes a test data set and divide it into three different data sets namely Training sample, validation sample and testing sample. A user has control in splitting up the data set as they see fit and specifying what percentage of data is to be used for setting up the model, validating it and testing the outcome. It creates a decision tree which is essentially a classification and regression tree. The training data set is split in to two parts at each stage, it tries to reduce the Gini coefficient. SPSS uses a similar technique in providing and preparing predictive models and is relatively fast in training and preparing model. XL Miner being in educational mode at the time of paper writing does not have much information for meaningful comparison with other packages and tools. However miner has capabilities to split the data into training, validation and test samples automatically

# ASSOCIATION ANALYSIS

Association analysis gives us an idea how different factors are associated with each other. also known as market basket analysis determines what products and services are purchased together by consumers in a retail setting in its most typical application. Rules describes patterns in dataset. For example, the analysis may reveal A and B are purchased item C is also purchased, where A and B are antecedents and C is the consequent. The output includes three quantities that measure the degree of uncertainty included with it. Support which is the probability that a randomly selected set of transactions from database expressed in percentage. Confidence is the conditional probability also expressed in percentage. Lift helps in measuring the improvement in probability .

Application analysis not only used in marketing and sales but also by insurance companies and online businesses. All three of them offered association analysis procedure and are quite similar in ease of user input data and output. User Friendly XL Miner offer the familiar MS Excel format SAS EM and Clementine offer a graphical user interface to run the procedure. XL Miner can handle two input data formats binary matrix formats and item list format, SAS EM and Clementine can only handle the item list input format. All three packages calculate support and confidence, SAS EM and XL Miner also calculate lift whereas Clementine output does not include lift.

# Conclusion

Some of the organizations using SPSS for their analytics work are International Business Machines (IBM), California State University-Stanislaus, National Opinion Research Center, Kelton Global, University of Illinois at Chicago etc. [6] While companies like Accenture, ICICI Bank, ICS consultancy Services use SAS for their analytical work load. XL Miner is used by renowned organizations like Northrup Grumman National Institutes of Health, NASA Centers for Disease Control, Monsanto, ExxonMobil, US Army, FDA, National Inst. of Standards & Technology, Fidelity Investments, Blackstone Group and more [5].

Of all the packages reviewed SAS provides the complete set of tools for data mining and analysis, while XL Miner gives us decent tools at a relatively low cost and setup complexities. For our term project we are would not be using any of these tools and will be using SQL Server and Microsoft business intelligence tools.

|  |  |
| --- | --- |
| [1] | D. Haughton, J. Deichmann, A. Eshghi, S. Sayek, N. Teebagy and H. Topi, "A Review of Software Packages for Data Mining," *The American Statistician,* pp. 290-309, 2003. |
| [2] | "What is data mining," SAS Institute inc., [Online]. Available: https://www.sas.com/en\_us/insights/analytics/data-mining.html. |
| [3] | "IBM SPSS Modeler Pricing," G2 Crowd, [Online]. Available: https://www.g2crowd.com/products/ibm-spss-modeler/pricing. |
| [4] | SAS, [Online]. Available: https://www.sas.com/store/products-solutions/cSoftware-p1.html. |
| [5] | "Solver," Solver, [Online]. Available: http://www.solver.com/xlminer-data-mining?gclid=CjwKEAjws-LKBRDCk9v6\_cnBgjISJAADkzXeCFXtEkTw-jT7O6vO\_I3We1CrBEVKAWQwmaG4SxWgMhoC\_tjw\_wcB. |
| [6] | "Descriptive Statistics," Investopedia, [Online]. Available: http://www.investopedia.com/terms/d/descriptive\_statistics.asp. |
| [7] | "List of companies using ibm-spss," i data labs, [Online]. Available: https://idatalabs.com/tech/products/ibm-spss. |
| [8] | "What is predictive modeling," Tech Target, [Online]. Available: http://searchdatamanagement.techtarget.com/definition/predictive-modeling. |