Many companies are gathering terabytes, petabytes, or exabytes of data. They are storing data in computer systems across networks, utilizing cloud computing, distributed processing, and data storage systems like Hadoop and IBM Netezza.

Consider integrated software environments such as IBM SPSS Modeler, SAS Enterprise Miner, Rattle for R, and Revolution R. What special features do these systems offer?

Each software environment listed above offers unique features, of which I will describe below:

IBM SPSS Modeler – This system allows users to build predictive models without needing programming background. I would consider this system to be closed box or black box in that it comes prepackaged with algorithms ready to be deployed at the click of mouse. In addition, this package also comes with Text and Entity analytics (IBM.com)

SAS Enterprise Miner – I see this system as being diverse across multiple business needs. For example it can accommodate both predictive and prescriptive analytics. This system will require that one can program (SAS.com).

Rattle for R – On rattle.togaware.com, one review stated that Rattle is similar to the ribbon interface for Microsoft. From reading the website, this tool is an entry level platform for graphics and modeling in the R environment.

Revolution R – This system is similar to Rattle in that it is for new comers to the R language, but it is also performs faster and is more reliable. The benefit of both Rattle and Revolution is that they are open source and continued to be expanded on by a dedicated community.

What are the possibilities for using a combination of programming tools to create a system for predictive analytics? What would be the advantages and disadvantages of such a system?

In regard to creating combination of programming tools to create a system for predictive analytics my pros and cons are listed below:

Pros – Multiple tools allow one to maximize time and efficiency based on the assumption that with more tools one can find the optimized solution.

Multiple systems allow for model validation such that confidence is built throughout the data process.

Cons – In my opinion, it would take a lot of time and money to learn all of these systems to be proficient enough to utilize the synergy.

In order to learn all of the systems above, one would naturally loose expertise in business application which can be dangerous in workplace setting.

References

Model Development and Deployment, Data Mining Software, SASÂ® Enterprise Minerâ„¢ | SAS. (n.d.). *SAS | Business Analytics and Business Intelligence Software*. Retrieved September 22, 2013, from http://www.sas.com/technologies/analytics/datamining/miner/

Revolution R Community Optimized Open Source R Programming Language. (n.d.). *Big Data Predictive Analytics with Enterprise R - Revolution Analytics*. Retrieved September 22, 2013, from http://www.revolutionanalytics.com/products/revolution-r.php

SPSS Modeler. (n.d.). *IBM*. Retrieved September 22, 2013, from www-01.ibm.com/software/analytics/spss/products/modeler/features.html

Togaware: Rattle: A Graphical User Interface for Data Mining using R. (n.d.). *Togaware: Rattle: A Graphical User Interface for Data Mining using R*. Retrieved September 22, 2013, from http://rattle.togaware.com/

*APA formatting by BibMe.org.*

What special challenges do predictive analysts face in working with extremely large data sets?

On a daily basis, I run into the challenges of working with extremely large data sets. At Target Corporation, I am often pulling down data from tables that have billions of rows and thirty plus columns. The challenges I face are the following:

Nomenclature – Our massive tables are stored in data warehouses from different vendors such as IBM, Teradata, and Host. Each vendor has different software to access the data warehouse like DB2, which comes with different coding nuances. Every couple of months our team is having to learn new coding differences such that best coding practices are utilized across the different systems.

Time – The proverbial bottleneck for analysis boils down to running code across multiple servers. Statistically I do not need to reference all the rows on the tables I am working with, rather knowing how the DBA’s have structured flag columns that reference 2% and 10% of the actual data is key for getting work done. The caveat for utilizing the flags is to know when to use just 2% of the data and when to pull more down.

What is your view of the software requirements for dealing with *Big Data*?

From what I have seen at US Bank and Target Corporation, *Big Data* is becoming more common as analytics drive decision making. Thus, robust software requirements are a *must* for an organization to make money from its data analysts. This translates to software that can do both heavy lifting (computing) and heavy learning. In addition, I have continued to see the need that software must have easy resolve capacity for when tables break or technical issues.

As a user of statistical software, what features do you regard as essential?

Before delving into my essential features, one should note that I am a rather new comer to statistical software and my preferences follow suit.

Adaptable: In my current software, work environment, I use different types of data in order to meet a client’s analytical needs. Much of my time needs to be focused on delivering analytics rather than focusing on the nuances of uploading and downloading datasets to my statistical software.

Intuitive: A synonym for intuitive is natural, and from my experience data analysis is often a multi-step iterative process. Having an intuitive statistical software package, including a clear log for interpreting errors, is the difference between a smooth analytical process or a situation that is overly painful and

frustrating.

Data Visualization: Communication is the cornerstone for working with teammates, managers, and clients. I have found that a well-developed data visualization will tell the story of the data across many different departments in an organization. In my current role, we use SAS and Tableau which turns out to be a rather complicated process. After executing the proper analytical techniques, I have to export the data into Tableau and spend time creating visualizations. I greatly value and desire a statistical software package that can deliver both top notch analytics as well as visualizations.