Data Mining
for Process Improvement

Paul Below, Quantitative Software Management, Inc. (QSM)

Introduction

What do you do if you want to create an estimate and you have 100 candidate variables to use in your estimating model?

This is also a common question for CMMI® high maturity organizations that need to create process performance models. According to SEI, process performance models are:

"A description of relationships among attributes of a process and its work products that is developed from historical process-performance data and calibrated using collected process and product or service measures from the project and that are used to predict results by following a process."

High maturity organizations typically use process performance models for operational purposes such as project monitoring, project planning, and to identify and evaluate improvement opportunities. They typically are used to predict many output variables including defects, test effectiveness, cost schedule and duration, requirements volatility, customer satisfaction, and work product size.[1]

Data mining techniques can be used to filter many variables to a vital few to build or improve predictive models. Specific examples are provided in four categories: classification, regression, clustering, and association.

When creating an estimating model or a process performance model, the primary challenge is how to start. Regardless of the variable being estimated (e.g., effort, cost, duration, quality, staff, productivity, risk, size), there are many factors that influence the actual value and many more that could be influential.

The existence of one or more large datasets of historical data could be viewed as both a blessing and a curse. The existence and accessibility of the data is necessary for prediction, but traditional analysis techniques do not provide us with optimum methods for identifying key independent (predictor) variables from a large pool of variables.

Data mining techniques can be used to help thin out the forest so that we can examine the important trees. Hopefully, this article will encourage you to learn more about data mining, try some of the techniques on your own data, and see if you can identify some key factors that you can control or use to build a predictive model.

What Is Data Mining?

There are many books on data mining, and each one has a slightly different definition. The definitions commonly refer to the exploration of very large databases through the use of specialized tools and a process. The purpose of the data mining is to extract useful knowledge from the data, and to put that knowledge to beneficial use.

Data mining can be viewed as an extension of statistical analysis techniques used for exploratory analysis, incorporating new techniques and increased computer power. A few sources are listed in the resources section that provide details on data mining.

There are a number of myths that have grown up regarding the use of data mining techniques. Data mining is useful but not a magic box that spits out solutions to problems no one knew existed. Still required for success:
- business domain knowledge
- the collection and preparation of good data
- data analysis skills
- the right questions to ask

Techniques for cleansing data, measuring the quality of data, and dealing with missing data are topics that are outside the scope of this article.

Researchers have created a number of new data mining algorithms and tools in recent years, and each has theoretical advantages and avid proponents. However, for the purpose of getting started with estimate model creation, tool selection is not critical. The comparative theoretical advantages and disadvantages of the techniques and tools is not important to our purpose of identification of key factors. The practical advice is to try as many different techniques as possible, as the difficult time-consuming task is data preparation. Refer to a list of tools in the References section.

Model Creation Challenges

People love to interpret noise. Regardless of what the data shows, the audience will offer theories to explain the causes for what is observed. If a graph shows that performance has improved, someone will offer an explanation for why that happened. If you tell the audience that the graph was upside down, and performance has actually decreased, just as quickly someone will propose a reason for why that happened.

Figure 1 is an image of random noise. If you stare at it long enough, you will start to see some patterns. People are pretty good at pattern recognition, even if no pattern actually exists. That is one reason why statistical quality control, data mining, and hypothesis testing are useful—to help us see whether the patterns we think we see are real or whether they could be explained by randomness alone. Another reason is to help us find patterns that are real but are difficult to see.

Figure 1: Random Noise
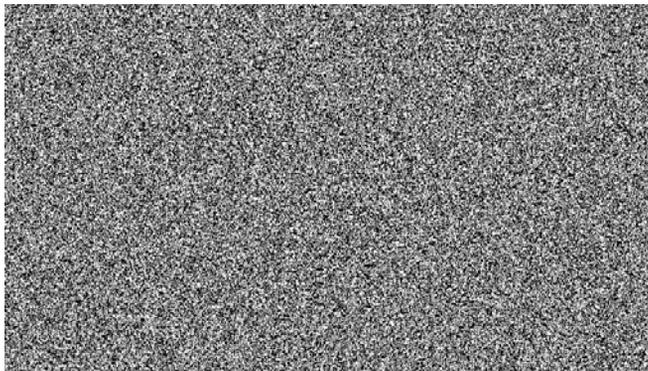
Exploratory analysis, including data mining, utilizes existing data that has already been collected. There are challenges with using such data, including:

• The databases already exist and almost always were created without considering analytical needs.

• Databases generally are built by committees, or have evolved from older systems through multiple stages. The variables stored include items that were used long ago as well as fields that someone thought might be useful someday, mixed in with data that are currently necessary. Many of the fields have values that are hard to decipher, or were used inconsistently by different populations of users.

• The structure of the data is often bad or the keys are not appropriate, making data extraction difficult.

Regardless of the data mining tools used, data extraction and validation is a major undertaking.

Once the data is extracted and placed in a readable format, the analyst is faced with dozens of input variables. Which of those variables should be used in the model?

It is common for our variables to exhibit colinearity. Colinearity is when the variables are highly correlated with each other. In practical terms this means that those variables are measuring the same or similar things. Dumping all of these variables into a regression equation is not a way to receive a useful output.

Data mining can help us thin out the forest so that we can see the most important trees. Many of the data mining techniques can be used to identify independent variables that are influential in predicting the desired result variable. Success will depend more on the mining process than on the specific tools used.

Data Mining Models

"Statisticians, like artists, have the bad habit of falling in love with their models."

- George Box

Data mining can aid in hypothesis testing as well as exploratory analysis.

There are many pure data mining products on the market, but they are typically very expensive. Some of the common techniques, however, are supported by basic statistical analysis tools which are much less costly. These techniques include all of the examples provided in this document. Examples of statistical analysis tools that support some or all of these functions are listed in the References section.

Data mining models can be placed into four categories as described in this table:

Table 1: Data Mining Models

Category	Description	Purpose	Primary Data Type
Classification	Split the data to form homogenous subsets	Predict response variable	Discrete is best
Regression	Best fit to estimating model	Predict response variable	Continuous (ratio or interval)
Clustering	Group cases that are similar based on selected variables	Identify homogeneous groups of cases	Any
Association	Group variables that are similar	Determine co-linearity, identify factors that explain correlations	Ratio or interval (not categorical)

We will now look at an example from each of the four categories.

Classification Example

One classification technique is a tree. In a tree, the data mining tool begins with a pool of all cases and then gradually divides and subdivides them based on selected variables.
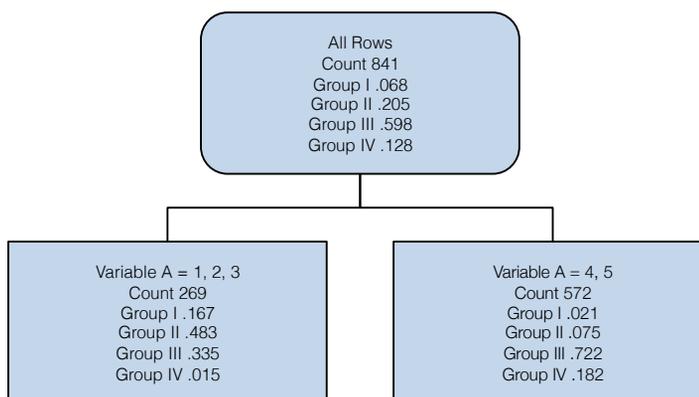
The tool can continue branching and branching until each subgroup contains very few (maybe as few as one) cases. This is called overfitting, and the solution to this problem is to stop the tool before it goes that far.

For our purposes, the tree is used to identify the key variables. In other words, which variables does the algorithm select first? Which does it pick second or third? These are good candidate variables to be used in an estimating model, since the tree selected them as the major factors.

In Table 2, we see an example that started with a data set of 841 cases, taken from a database of client information. Prior to running the tree, each of the 841 clients was assigned to one of four groups. The assignments were made based on information about customer satisfaction. The goal of the analysis was to see if there were key factors that could be used to predict which group a client would fall in. This prediction would then be used to identify clients that were likely to become less satisfied in the future, and determine actions that could be taken to improve client satisfaction.

In the top box of the tree, each group is listed with the fraction of the cases. So, for example, Group I contains 6.8% of the 841 cases. The total for the four groups will be approximately equal to 1 (100%) allowing for round off.

Table 2: Tree Example

All Rows
Count 841
Group I .068
Group II .205
Group III .598
Group IV .128

Variable A = 1, 2, 3
Count 269
Group I .167
Group II .483
Group III .335
Group IV .015

Variable A = 4, 5
Count 572
Group I .021
Group II .075
Group III .722
Group IV .182

The tree algorithm examined all of the variables and selected Variable A to be the first branch. Variable A has possible integer values from 1 to 5. As we can see, the algorithm put the cases where Variable A is equal to 1, 2, or 3 in the left branch and those with Variable A equal to 4 or 5 in the right branch.

The left branch has 269 cases, including most of the cases in Groups I and II (the 269 cases are composed of 16.7% Group I and 48.3% Group II, compared to the right branch which is composed of only 2.1% Group I and 7.5% Group II). The right branch ended up with 572 cases, including most of the cases in Groups III and IV.

Variable A by itself is not a sufficient predictor to use as a predictive model. However, the tree is telling us that Variable A is one important factor.

The tree would have additional branches, but Table 2 is sufficient to aid in explaining how the tree is used.

Regression and Correlation Examples

The data used in the remaining examples came from industry data. It is based on a sample of 193 projects extracted from a corporate database.

The output in the examples is for illustrative purposes and should not be used to reach conclusions about performance of specific software projects.

Stepwise regression is a type of multivariate regression in which variables are entered into the model one by one, and meanwhile variables are tested for removal. It can be a good model to use when supposedly independent variables are correlated. Stepwise regression is one of the techniques that can help thin out the forest and find important predictive factors.

Table 3 is a summary output of a stepwise regression that went through nine steps to build the best model. It was created in SPSS, although other statistical packages produce similar results. The dependent variable being predicted was errors detected prior to deployment. The stepwise regression selected nine variables that fit the threshold for inclusion, while excluding 20 other variables (not listed).

Table 3: Regression Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
9	.840	.706	.691	330.332

Model		Sum of Squares	df	Mean Square	F	Sig.
9	Regression	47883321.563	9	5320369.063	48.757	.000
	Residual	19968796.365	183	109119.106		
	Total	67852117.927	192			

Predictors: (Constant), Effective SLOC, Life Duration (Months), MB Time Overrun %, MB Effort (MM), Life Peak Staff (People), Data Complexity, MBI, MB Effort %, Mgmt Eff.
Dependent Variable: Errors (SysInt-Del)

The nine variables selected by the stepwise regression were, in the order the tool selected them: effective source lines of code; project life cycle duration in months; percent of duration overrun of Main Build (design through deploy); Main Build man months of effort; peak staff; data complexity; Putnam's Manpower Buildup Index; percent of effort expended in Main Build;

and management effectiveness. Note that two of these nine variables (data complexity and management effectiveness) are qualitative, scored on a scale of one to 10 where five is average and 10 is high.

The first number to look at in Table 3 is the Sig (significance) in the rightmost column. The most commonly used significance threshold is .05, which means that the variable or model would be significant at the 95% level. In the example, the value .000 means that we have less than a one in a thousand chance of being fooled by random variation into thinking this model is significant.

Although all nine variables selected are clearly significant, the overall model created has an adjusted R square of .691, which means that these nine variables taken together are explaining about 69% of the variation in errors found. This may not be the best model to use for estimating, but it is important to look at each of the nine variables if the intent is to create an estimating model or if we need to reduce the number of errors found in the future.

The coefficients of the stepwise regression formula are displayed in Table 4. Each variable is listed next to the coefficient B, which is the multiplier in the linear equation.

Table 4: Regression Coefficients

Coefficients:
Dependent Variable: Errors (SysInt-Del)

Variable	Unstandardized Coefficients		Sig.	95% Confidence Interval for B	
	B	Std. Error		Lower Bound	Upper Bound
(Constant)	-580.411	239.656	.016	-1053.255	-107.568
Effective SLOC	.001	.000	.000	.001	.001
Life Duration (Months)	27.633	5.832	.000	16.126	39.139
MB Time Overrun %	.026	.006	.000	.015	.037
MB Effort (MM)	1.535	.326	.000	.892	2.177
Life Peak Staff (People)	-7.438	1.905	.000	-11.197	-3.679
Data Complexity	66.840	18.269	.000	30.795	102.886
MBI	33.683	14.609	.022	4.859	62.507
MB Effort %	3.924	1.552	.012	.862	6.987
Mgmt Eff.	-50.012	22.775	.029	-94.948	-5.076

The equation that yielded the adjusted R square of .691 is:
Errors = -580 +(.001*ESLOC)+(27.6*Duration)+(.026*overrun%)+(1.5*MB Effort)-(7.4*peak staff)+(66+data complexity)+(33.68*MBI)+(3.9*MB effort %)-(50*Mgmt Eff)

The factors in the equation can be determined from reading the numbers in the B column.

A negative number means a negative correlation. One counterintuitive result of this example is the coefficient for peak staff. The negative coefficient means in this model the larger the peak staff the smaller the number of errors detected. This type of result is why it is necessary to evaluate the data in more depth and do additional analysis before using the model. Sometimes, negative correlations are expected. For example, management effectiveness has a negative coefficient meaning that a higher effectiveness results in a lower number of errors.

The two rightmost columns, the 95% confidence intervals, are useful as an indication of the uncertainty in the coefficients. The lower and upper bound for any variable should not straddle zero. If it did, that would be an indication that we lack confidence in the factor B. Another method is to compare the value of the standard error to the value of the coefficient; ideally the standard

error should be much smaller than the coefficient B. Also, the Sig should be small, ideally less than .05.

In addition to regression, correlation can be used to identify candidate important variables. This can be done by selecting the dependent variable first for the correlation and then the list of independent variables. There are different types of correlation that can be used. For ratio data, Pearson correlation can be used. For ordinal data, Kendall's Tau-B will work. For nominal (categorical) data, a chi square test can be used on a crosstab (two-way table) to determine significance.

It is important to note that these tests will determine linear correlations. Sometimes correlations exist but are nonlinear. One technique for exploring those relationships is transformation, which is not discussed in this paper.

Clustering Example

Cluster techniques detect groupings in the data. We can use this technique as a start on summarization and segmentation of the data for further analysis.

Two common methods for clustering are K-Means and hierarchical. K-Means iteratively moves from an initial set of cluster centers to a final set of centers. Each observation is assigned to the cluster with the nearest mean. Hierarchical clustering finds the pair of objects that most resemble each other, then iteratively adds objects until they are all in one cluster. The results from each stage are typically saved and displayed numerically or graphically as a hierarchy of clusters with subclusters.

Table 5 is the output of a K-Means example run from the sample with the output constrained to create exactly three clusters.

The tool placed the largest projects in the first two clusters. These projects had more errors, more staff, and higher productivity than the third cluster. One difference between the first two clusters is that the projects in the second cluster tended to have poor estimates of effort.

Table 5: Cluster Example

Final Cluster Centers

	Cluster		
	1	**2**	**3**
Project Count	5	22	166
Life Effort (MM)	750.7	617.8	89.1
Errors (SysInt-Del)	1898	1030	186
Errors First Month	138	117	8
Total FP	37167	26533	2648
Effective SLOC	1272194	298791	26444
Life Duration (Months)	21.3	18.4	9.3
Life Peak Staff (People)	56.5	61.1	15.4
Life Avg Staff (People)	23.8	26.5	7.1
MB Eff Overrun %	.0	62.0	45.8
SLOC/MB MM	2384.5	1606.4	910.9
Putnam's PI	24.4	21.5	14.1

We may want to stratify the projects into groups based on the above distinctions prior to conducting additional analysis. This may result in the need for more than one estimating model, or more than one process improvement project.

Association Example

Association examines correlations between large numbers of quantitative variables by grouping the variables into factors. Each of the resulting factors can be interpreted by reviewing the meaning of the variables that were assigned to each factor. One benefit of association is that many variables can be summarized by just a few factors.

In the following example using actual data, Principal Components analysis was used to extract four components. The Scree Plot in Figure 2 was used to determine the number of components to use. The higher the Eigenvalue, the more important the component is in explaining the associations. Selection of the number of components to use is somewhat arbitrary, but should be a point at which the Eigenvalues decline steeply (such as between components 2 and 3, or between 4 and 5). It turned out in this example that the first four components account for roughly half of the variation in the data set (included in other output from the principal components tool, not shown here), making four a reasonable choice.

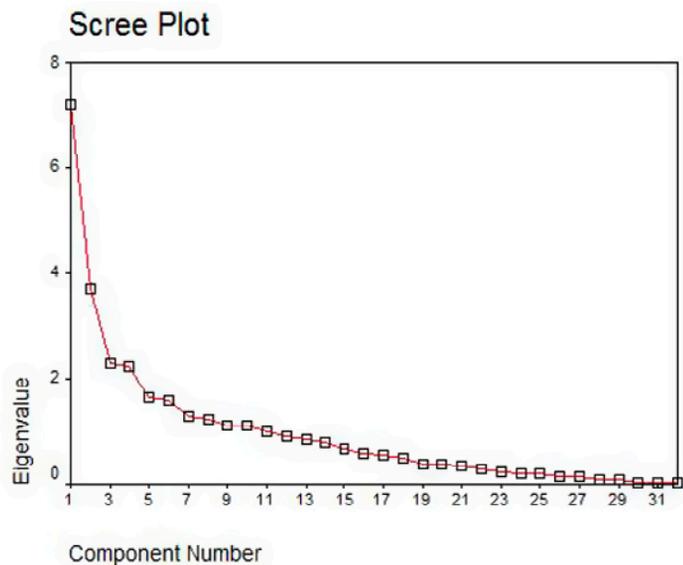*Figure 2: Scree Plot for Association example*

Table 6 displays variables with the most significant output for each component. The important numbers in the table are those with relatively large absolute values and have been shaded for easy reference.

• Component 1 is composed of a market basket of variables related to effort and size (the variables aligned with the shaded numbers in component 1).

• Component 2 grouped variables related to the development team: knowledge, turnover, and skill.

• Component 3 isolated the Manpower Buildup Index, which is the speed at which staff is added to a project.

• Component 4 linked the percent of effort expended in functional requirements to the percent expended in the Main Build (design through deploy).

Variables that are seen to be related should be combined (or one should be chosen as the representative) as an input variable when creating prediction models or identifying root causes.

Table 6: Association example output

	Component			
	1	**2**	**3**	**4**
Life Effort (MM)	.920	-.152	.196	-.006
Effective SLOC	.652	.111	-.475	.106
Life Duration (Months)	.658	-.198	-.429	.066
Life Peak Staff (People)	.865	-.115	.338	-.137
Life Avg Staff (People)	.823	-.157	.381	-.156
FUNC Effort (MM)	.880	-.169	.151	.098
MB Effort (MM)	.925	-.160	.065	-.122
Func Effort %	-.241	.088	.236	.719
MB Effort %	-.059	-.072	-.247	-.765
Knowledge	.186	.770	.161	-.076
Staff Turnover	.083	-.717	.049	.110
Dev Team Skill	.133	.746	.029	-.225
MBI	-.006	-.011	.640	-.200

Summary

Once data has been collected and validated, the hardest work is behind you. Any data mining tools that are available to the researcher can be used relatively quickly on clean data. These data mining techniques should be used to filter an overwhelming set of many variables down to a vital few predictors of a key output (for example, quality).

Determination of the vital few is a key component of process improvement (such as Six Sigma projects) activities as well as prediction. With those key drivers or influencers of quality in hand, improvements can be designed and implemented with fewer iterations, effort, or time.

In addition to process improvement activities, we use the "vital few" to build error prediction models, and then use the models to tune parametric project estimates for specific clients. The project estimate and plan is thereby not only an estimate of duration and cost to complete construction, but also includes the prediction of when the system will be ready for prime time.◈

Disclaimer:
® CMMI is registered in the U.S. Patent and Trademark Office by Carnegie Mellon University.

ABOUT THE AUTHOR

Paul Below has over 25 years experience in the subjects of measurement technology, statistical analysis, estimating and forecasting, Lean Six Sigma, and data mining. He has provided innovative engineering solutions as well as instruction and mentoring internationally in support of multiple industries. He serves as services consultant for Quantitative Software Management (QSM) where he provides clients with statistical analysis of operational performance, helping strengthen competitive position through process improvement and predictability.

Paul is a Certified Software Quality Analyst, and a past Certified Function Point Specialist. He is a Six Sigma Black Belt. He has been a course developer and instructor for Estimating, Lean Six Sigma, Metrics Analysis, Function Point Analysis, as well as statistics in the Masters of Software Engineering program at Seattle University. He is a member of the IEEE Computer Society, the American Statistical Association, the American Society for Quality, the Seattle Area Software Quality Assurance Group and has served on the Management Reporting Committee of the International Function Points User Group as well as CMMI high maturity assessment teams. He has one US patent and two patents pending.

paul.below@qsm.com
Quantitative Software Management, Inc. (QSM)
http://www.qsm.com/

RESOURCES

Data Mining Websites:

<http://www.twocrows.com> <http://www.kdnuggets.com> <http://www.datamininglab.com>

Data Mining Tools:

• Statistical tools that each do some data mining techniques include: SPSS; SAS; JMP; SPlus; Minitab

• Specialized data mining tools include Salford Systems CART and MARS; SAS Enterprise Miner; PASW Modeler (SPSS); Insightful Miner (SPlus)

Books:

• *Introduction to Data Mining*, by Pang-Ning Tan, et al, Addison-Wesley, 2006.

• *Principles of Data Mining*, by David Hand, Keikki Mannila and Padhraic Smyth, MIT Press, 2001.

• *Data Mining – Concepts, Models, Methods and Algorithms,* by Mehmed Kantardzic, John Wiley and Sons, 2003.

• *Data Mining: Opportunities and Challenges,* by John Wang, IDEA Group, 2003.

• *Use and Organizational Effects of Measurement and Analysis in High Maturity Organizations: Results from the 2008 SEI State of Measurement and Analysis Practice Surveys,* by Dennis Goldenson, James McCurley, Robert W. Stoddard, CMU/SEI-2008-TR-024.

NOTES

1. CMU/SEI-2008-TR-024