Paper-071-31

Investigating Open Source Project Success: A Data Mining Approach to Model Formulation, Validation and Testing

Uzma Raja and Marietta J. Tretter
Texas A&M University, College Station, TX

ABSTRACT

This paper demonstrates the use of Data Mining (DM) techniques in exploratory research. A robust model for identifying the factors that explain the success of Open Source Software (OSS) projects is created, validated and tested. The predictive modeling techniques of Logistic Regression (LR), Decision Trees (DT) and Neural Networks (NN) are used together in this analysis. Using Text Mining results in the predictive modeling process strengthens the model. SAS® Enterprise Miner and SAS® Text Miner are used in this research.

INTRODUCTION

The term OSS refers to software distributed by license that conforms to the Open Source Initiative (Feller et al. 2002). The most common of these licenses are GNU General Public License (GPL) and Berkley Software Distribution (BSD). OSS projects are typically developed through online communities for software development (e.g. SourceForge.net, Freashmeat.org and Tigris.net) which offer project-hosting resources. These communities serve as a “Bazaar” where developers and end users can come together and find suitable matches to their skills and requirements. Users can view, update and change the project source code. They can also detect and report bugs, request new features, and at times, contribute to the software source code.

According to a recent survey (Cearly et al. 2005), the use of OSS projects in the business community is increasing. There is a need for models that can be used to identify the projects that are more successful than others. This will help the users and the corporations, which intend to use OSS projects, to make the right choice of projects. The existing project performance evaluation models were developed for commercial software systems (CSS) and cannot be used for OSS. This is because most of the performance measures of CSS projects (e.g. cost, schedules) hold no meaning in OSS domain. Therefore, exploratory research has to be conducted to identify the factors that are critical in the success of these projects.

Keeping with the philosophy of free sharing, OSS communities grant access the project source code, and other artifacts (e.g. e-mail communications, bug reports, number of downloads and information on developers) of the projects. The lack of any tested models and theories for OSS projects and the availability of rich datasets make this domain a good candidate for exploratory research using DM. Exploratory research can be used to discover new phenomenon and to develop new theories. In past, access to software lifecycle data was a challenge and researchers had to rely on limited datasets to develop and test models. However availability of OSS data archives presents a unique opportunity for the research community to develop models based on transactional data and to extract knowledge from it.

In this research a performance evaluation model for OSS projects in their development phase is formulated. This model identifies some critical factors that contribute to the success of OSS projects. The identification of the factors is very important to the practitioner and research community. The development teams can better monitor the performance of their projects and adjust the input variables to achieve the desired outcome. For businesses that intend to adopt OSS projects, this provides them with the ability to make better decisions regarding adoption. For the research community the model provides a deeper understanding of the phenomenon of OSS development and a better model to predict software project outcomes. It also identifies some new factors that have not been used in prior research.

DM techniques can be used for model formulation, testing and validation (Brachman et al. 1996). The three DM predictive modeling techniques of; LR, DT and NN are used in this research. Each of these techniques has its strengths and weaknesses. The best model is selected on the basis of fit statistics and ease of explanation of the relationships. Knowledge from Textual documents is also used in conjunction with the DM techniques to improve the model.

PROBLEM DEFINITION

OSS projects are usually available for free download and can be modified at the user end. But the cost of switching to OSS applications can be significant if a business or end user invests time, effort and resources in adopting an OSS project that fails. Therefore there is a need for models that can predict the outcome of OSS projects and that can identify the factors that contribute to the success of these projects. The available models to evaluate the outcomes of software projects were developed for CSS projects. The performance measures used in these models (e.g cost, schedule and conformance to requirements), hold no meaning in OSS domain.

While creating a new model for OSS project success, the first challenge is to identify the critical and relevant variables from the data warehouse. Traditionally, product, process and resource characteristics have been used in models of performance evaluation for CSS projects. Product characteristics refer to the attributes of the project that are product specific e.g. reliability, maintainability etc. Process characteristics refer to the attributes of the process through which the product is developed e.g. use of configuration management techniques, use of project management tools etc. The resource characteristics refer to the attributes of the resources involved in the development of the project e.g. the size of the development team.

In OSS, besides the above-mentioned characteristics, the role of end-users can be very important in the success of the project. Since there is no formal project requirements elicitation, end-user involvement in the process of development can be critical. Especially in projects that are end-user applications. End-users can detect and report errors, submit fixes, submit support requests and suggestions regarding the project. For projects that are primarily for the use of the developer community, this might not be very significant since the development team themselves can play the role of user. Considering this situation in OSS, the end-user characteristics will also be considered in the model for project performance evaluation.

The next challenge is to identify a rich dataset for the use in the DM process. OSS communities maintain transactional datasets on the projects they host. Recently, SourceForge.net has started to share this data with the research community so that these rich datasets can be used in research on software systems development. The data warehouse for SourceForge.net contains over 100 tables and over 1000 variables for over 100,000 projects. It contains data artifacts on all aspects of project development and maintenance. The Sourceforge.net data archives starting November 1999 through May 2005 are being used for this research (Madey 2005). Table 1 describes some variables identified for use in the analysis.

	Attribute	**Measure**
Product Characteristics	Reliability	Time Taken to remove an error
	Functionality	Number of modules added to the project
	Use of Config Management	A binary variable (No = 0, Yes = 1)
Process Characteristics	Use of Communication tools	A binary variable (No = 0, Yes = 1)
	Team Size	Count of the number of developers of the project
Resource Characteristics	User Community Size	Count of the distinct members posting messages
	Usage	Number of downloads of the project
User Characteristics	User Activity	Count of the number of messages posted online
	User Contributions	Bugs reported by the end users of the project
Domain Characteristics	Audience	Developers = 0, End-users = 1
	Domain	Type of the project
Outcome	Growth	Number of new versions released by the project
	Evolution	Transition of project through development phases

Table 1: Some important variables used in the analysis

MODEL BUILDING USING SAS® ENTERPRISE MINER

Availability of large amounts of transactional data and need for exploratory research, makes OSS an ideal candidate for using DM techniques. DM allows the flexibility of using a large number of variables and extracting potentially useful models from them. The three main types of predictive modeling resources used in DM are LR, DT and NN. All these three techniques are used in this research. Results obtained from each technique are carefully analyzed to improve the subsequent runs. DT can be used to identify any interaction terms, while the NN analysis can be used to

detect any non-linear relationships that may exist.

SAS® Enterprise Miner provides the computational power to analyze large datasets and also provides modeling techniques to create validate and test models from the data. The following sections describe the use of SAS® Enterprise Miner and SAS® Text Miner in formulating the model for OSS projects success. Figure 1, shows process flow of the analysis.

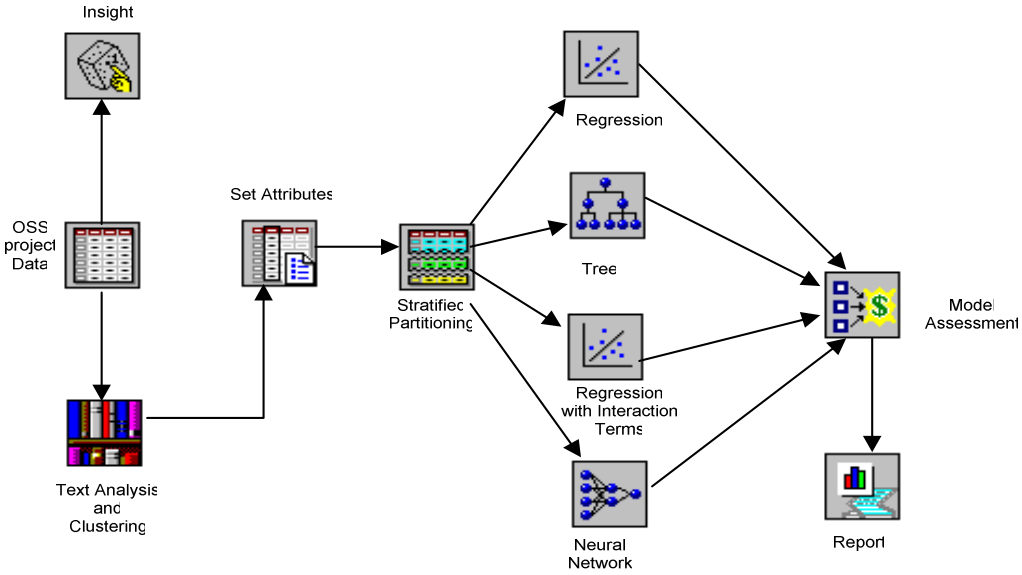

Figure 1: Process Flow of the Analysis

DATA SOURCE

The data source node is used to select the required data set and to set the status of the variables to be used in the analysis. The model role of the variable representing the project success is set to "Target". Redundant variables are set to status "reject", while the variables that are to be used in the analysis are set to the status "input". The data source node also provides the option to set the measurement scale of the variable. This node can be used to ensure that all variables are in the correct measurement scale. There were binary, interval and nominal variables. The data source menu also provides the descriptive statistics for the interval variables. The percentage of missing values, if any, can also be seen in the menu. After data cleaning, there were 4931 observations and 36 variables. The percentage of successful projects was 16.5% of the total projects. Figure 2, shows the frequency distribution of the target variable.

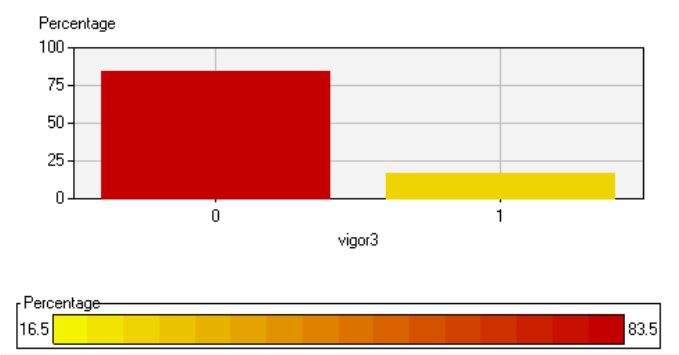

Figure 2: Frequency distribution of the target variable

INSIGHT NODE

Analyzing the descriptive statistics of the dataset is a very significant part of any model building process. The Insight node has several useful options that are critical in exploring large datasets. Distributions of the important variables can be viewed to get more information on the relationships. The pair-wise correlations of the variables give useful information on the relationship they might have with each other and with the dependent variable. The insight node is also used to test for multicollinearity. It provides very important statistical analysis e.g. Variation Inflation Factor (VIF) and Conditional Index (CI) to detect multicollinearity. For the OSS data, both the VIF and the CI were within the acceptable range, therefore presence of multicollinearity is ruled out. Presence of missing values can corrupt the results of the analysis. Insight node was used to detect the missing values, which can be either replaced by appropriate values, or the decision to remove the entire observation can be made.

TEXT ANALYSIS NODE

Projects used in this analysis, are diverse in size, application domain and audience. Besides various categorization variables, each project maintains at least a 200 word long description of the scope of the project. This information was used in the Text Analysis node to cluster the projects on the basis of this textual information. Textual data can provide additional information that is otherwise hard to capture from numeric variables.

Initially the default “stop” list was used to perform the analysis. The terms were later analyzed and refined to create a new “start” list. The automatic clustering option was selected, using the exception maximization method. The clusters were based on the singular value decomposition (SVD) dimensions. The term weighting method used was entropy.

The results of the analysis yielded six clusters. The resulting clusters were analyzed (e.g. the cluster 4 contains projects that are related file transfer applications). The resulting dataset was saved and the cluster_id was used in predictive modeling to improve the model performance. The set attribute node was then used to identify the target and the variables to be used in the analysis. Besides the initial set of variables, the cluster ID of the project was also used. A screen shot of the results window is shown in figure 3.

Text analyst can play a very important role in academic research. The use of textual data in quantitative analysis can significantly improve the accuracy and applicability of the results. The amount of textual data maintained by organizations and individuals is increasing. Therefore text analyst holds promise for academic research. It can provide a deeper understanding of the variables that are otherwise hard to quantify.

Description	group_id	evol	bugcnt	bugreportedbynu	countusr
<Emacs is a highly customizable text editor and application development system.	11	0	5	3	3
See Anteil. Apollo was a joint effort between VA Linux Systems and Keystone Pr	18	0	129	19	1
VA-Cluster Manager is a large scale server monitoring and management tool. Pro	19	0	8	5	1
The FreeWorld BBS is an attempt to create an easy to use, yet robust platform to	34	0	2	0	4
dents is a server implementation of the Internet's Domain Name System with a fo	63	0	7	4	1
pac is an ip accounting package for linux. It collects,summarizes and nicly displa	70	0	8	3	1
gchbkgrd is a program used for automatic and periodic change of the desktop's	82	0	1	0	.
Gnofin is a light-weight personal finance application for GNOME.	102	0	26	21	4
SCREEM is a GNOME website / tag based html editor (ie not WYSIWYG) which	142	0	223	159	6
Dialid is an intelligent link management tool originally named for its	179	0	25	13	2
acmemail is a multiuser IMAP/POP3 to Web gateway (or webmail program). It rea	186	0	28	7	8
GnuCash is a personal finance manager. A check-book like register GUI allows u	192	0	11	5	5

16,719 Terms — Display dropped terms — Display kept terms — Filter — Find Similar

Term	Freq	# Documents	Keep	Weight	Role
+ project	1520	1337	Y	0.229	Noun
+ write	1348	1287	Y	0.229	Verb
+ base	1313	1231	Y	0.235	Verb
+ application	1257	1115	Y	0.248	Noun
java	1268	1071	Y	0.254	Prop
+ tool	1054	961	Y	0.263	Noun
+ allow	996	943	Y	0.263	Verb
+ file	1188	932	Y	0.272	Noun
+ user	969	866	Y	0.275	Noun
+ library	954	850	Y	0.277	Noun
+ support	808	758	Y	0.287	Verb
+ web	842	720	Y	0.296	Noun

Clusters — Filter — Find Similar

#	Descriptive Terms	Freq	Percentage	RMS Std.
1	+ source, de, open source, + program, open	648	6%	0.1131220221
2	mysql, + base, + game, + file, + web	2188	21%	0.1248433739
3	windows, + driver, os, + support, + run	1405	13%	0.1238284858
4	+ server, + client, irc, + protocol, + write	743	7%	0.1174653117
5	java, + tool, + application, data, + language	3094	29%	0.1267024564
6	+ support, information, + design, + develop, + project	2534	24%	0.1272855508

Figure 3: Text Analysis output of the project description data

SAMPLING NODE

In exploratory research, it is very important to partition the data into train, validate and test samples. If the same data were to be used for model creation and validation, the resulting model would likely be biased to the sample and thus not acceptable. Therefore the original dataset is split into training, validation and testing datasets. This is done to ensure that a valid model is created that will be applicable to OSS projects in general. Initially, the training set is used to train or build the model. Once an acceptable training model is built, the validation set is used to evaluate the model. A comparison is made with specific diagnostics e.g. lift charts to check how well the training model holds for the validation dataset. Based on the fit of the model, there might be a need to further improve the model. There can be several iterations of re-training before a reasonable model is selected. Once a model is selected, the validation data can no longer be used to test the accuracy of this model. To create a robust model, the testing dataset was used to fit the model. The accuracy of the model on the test data gives a realistic estimate of the performance of the model for OSS projects in general. The original dataset was split into 40% build, 30% validate and 30% test splits for the analysis. The descriptive analysis of the dataset had revealed that the number of successful projects was low. Therefore stratified sampling was used.

REGRESSION NODE

The Regression node is one of the predictive modeling techniques used in this research. Since the target is a binary variable, LR was used. Stepwise method was used for variable selection. The training data was used to build the initial model. The model was then fine tuned using the validation data. The testing data was used to get the unbiased estimate of the generalization error of the model. The regression node result window provides various plots, statistics and estimates to analyze the model. The results identified the variables that were found significant in the analysis. It also provides the fit statistics of the model. For multiple runs, various diagnostic tests can be performed to select the best model.

Initially the results of LR were not as good as the ones from NN or DT. These initial results were used to identify interaction terms and higher order terms. A new LR node was added to the analysis. The interaction builder was used to create new interaction terms and the transform node was used to create new higher order terms (e.g. square of downloads). The subsequent runs indicated a significant improvement in the results obtained through the LR.

TREE NODE

The Tree node is used to perform the DT analyses of the data. SAS® Enterprise Miner Tree node offers various selection options for the datasets. The advantage of using DT is that the results can be explained in simple English rules. Figure 4 shows the misclassification rate of the target using the Tree node. As mentioned earlier, the initial results indicated that the performance of DT and NN was better than LR. DT results can be effectively used to identify interaction between the independent variables. In this analysis, the DT results indicated that there was an interaction effect between the end-user activity and the intended audience of the project. Therefore this effect was incorporated in later LR models to improve its performance.

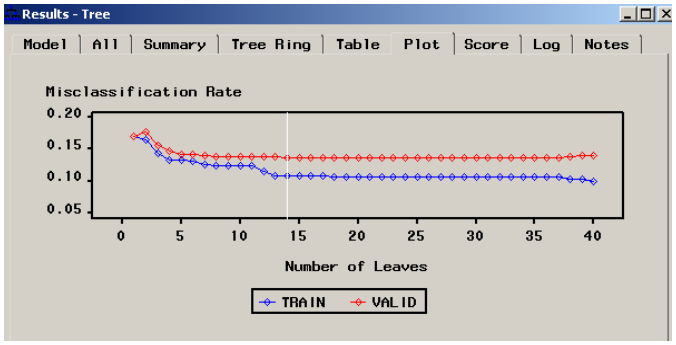

Figure 4: DT preliminary results

NEURAL NETWORK NODE

The NN node provides the ability to fit data into models with high predictive power. Various selections of the available options for hidden nodes are used before the best model is selected. The problem with the results of a NN is that it's very difficult to interpret the results. In exploratory research, it is very important to be able to identify the

factors that are significant in the model. In the initial analysis, the performance of the NN node was superior to LR and DT. This indicated that there could be presence of higher order effects. Therefore, after analyzing the results of NN, new higher order terms were added to the LR model. The results were considerably improved. This demonstrates that the three techniques can be used together to improve the performance of the final model.

ASSESSMENT NODE

The assessment node provides the ability to compare and assess multiple models and techniques. The assessment node was used to analyze the models. The fit statistics, misclassification rates, lift charts and ROC curves are all very useful in model assessment and selection. This node is very useful in exploratory research, since it provides a wide range of diagnostic features.

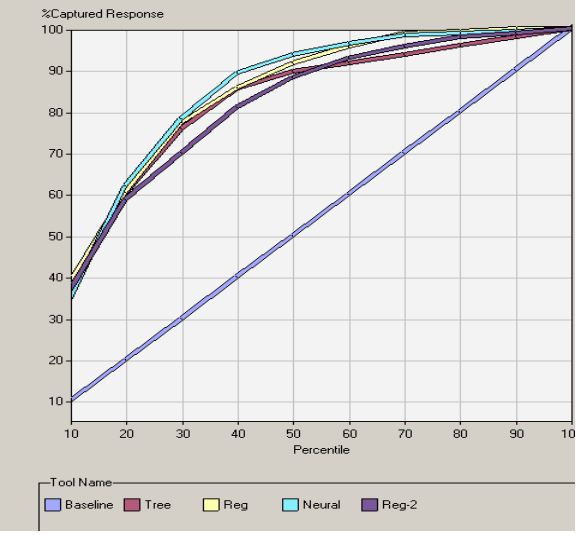

Figure 5: Lift Chart for the analysis

REPORT NODE

A very significant part of any research is the reporting of the results. The reporter node provides the analysis in a format that is useful in preparing scientific results and reports. Addition of some features required for reporting research results (e.g. VIF and CI for regression) would make it more suitable for research applications.

RESULTS AND CONCLUSION

The dataset was analyzed using LR, DT and NN nodes. The initial results from DT and NN were far superior to LR, based on diagnostic testing and fit statistics. Useful information regarding interaction effects and presence of higher order terms was extracted through analyzing the results of DT and NN. Therefore a new LR node which accounted for potential interaction and higher order terms was added. This improved the performance of LR analysis and resulted in a model with lowest AIC. This demonstrates that for academic research various techniques can be used simultaneously to improve the final model.

According to the preliminary findings of this research, the projects that were created before the year 2003 were less likely to succeed as compared to the more recent projects. One of the reasons can be that OSS movement is becoming more popular and the newer projects offer more promise to developers and the users compared to the older projects. This would also imply that with time, OSS teams are improving their project management process. Another important finding is, that the number of downloads are positively related to success. Projects that have more downloads are more likely to succeed. The number of bugs reported has a positive relationship to success. The number of bugs in CSS is typically reported to have a negative relationship to project success. But in case of OSS, this can be an indicator of the usage of the project rather than its quality. Therefore, the higher the number of bugs reported, implies that the software is being used and therefore has a positive relationship to success. The number of bugs open is an indicator of the inability of the project team to fix the bugs; therefore it has a negative impact on success. The team size has a positive impact on success, so the bigger the team size, the probability of success of the project increases. OSS projects also have the option to use a project manger or not. Use of project management methods has a positive impact on success of the project.

This research demonstrates the use of SAS® Enterprise Miner in academic research. It has the ability to analyze large datasets and to create and validate new models. SAS® Text Miner can be very useful in supplementing the predictive models by extracting useful knowledge from textual documents. This research also highlights the combined use of multiple predictive modeling techniques, to improve the performance of the models.

REFERENCES

Brachman, R.J., and Anand, T. "The Process of Knowledge Discovery in Databases: A human Centered Approach," in: *Advances in Knowledge Discovery And DataMining,* U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthurusamy (eds.), AAI Press/The MIT Press, Menlo Park CA, 1996.

Cearly, D.W., Fenn, J., and Plummer, d.C. "Gartner's Position on the Five Hottest IT Topics and Trends in 2005," G00125868.

Feller, J., and Fitzgerald, B. *Understanding Open Source Software Development* Addison-Wiley, London, 2002.

Krishnamurthy, S. "Cave or Community? An Empirical Examination of 100 Mature Open Source Projects," *First Monday* (7:6) 2002.

Madey, G. "SourceForge.net Research Data Archive," http://www.nd.edu/ ~oss/Data/data.html, 2005.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the authors at:

Uzma Raja
Department of Information and Operations Management
Mays Business School, Texas A&M University
College Station, TX, 77840
Work Phone : (979) 845-6995
Fax: (979) 845-5653
E-mail: uraja@mays.tamu.edu
Web: iops.tamu.edu/faculty/uraja