Improving Budgeting for the Green Building Certification Institute

BUDT733 – Data Mining for Business
Fall 2009

Submitted By:

Dean DiPietro
David Kuhla
Anuja Rathi
Jay Virgil
Paige Washington

Executive Summary

This report aims to help the Green Building Certification Institute (GBCI) better utilize data related to the Leadership in Energy and Environmental Design (LEED) Green Building Rating System. LEED is a voluntary certification program administered by GBCI, a 501c non-profit organization. The rating system is designed to evaluate building design, construction, and operation across an array of environmental performance criteria including site selection, energy and water use, material waste, and indoor environmental quality. GBCI administers certification only after conducting a thorough review of project documentation.

Our analysis sought to provide GBCI with an analysis of the registration to application process, seeking to determine if there were any factors that led projects to be more or less likely to submit an application and how quickly that action would occur. The most important part of the budget process at GBCI is forecasting the number of projects that will formally apply. This drives staff levels and resource allocation, a critically important measure for a non-profit organization operating under tight funding constraints. The organization would like to have some idea of what its budget will be for the next year, along with the necessary staff to accommodate projects seeking certification. Setting a budget too low means that current GBCI staff will face some strain in completing all necessary tasks, and setting a budget too high means wasting money and hiring un-needed staff.

Customers go through a multi-step process to achieve certification as shown in Appendix A. The present challenge is that customers are completing the initial registration step at a rapidly increasing pace, from only a few hundred registrations in 2005 to over 2,000 in 2008, and the trend into the future is for continued growth in demand. However, the follow-through on actually applying is low and the ratings systems have been changing quickly. Most importantly, more than 75% of GBCI's revenues are derived from the corresponding fee associated with application.

Our analysis included a rigorous data exploration process which looked at all the variables as potential predictors for application status. Data was provided by GBCI and covered all projects (about 30,000 records) registered from 2000 to October 2009. Utilizing data visualization tools such as Spotfire and Excel, we determined the most likely predictor variables. Due to the complexity of the task at hand, a regression analysis that would have provided GBCI with a continuous Y variable (i.e., a method to plug in known data about a project to provide an estimate time when it would be registered) was rejected—there were just too many projects that never moved to application to make such an analysis feasible. We then looked at categorical variables such as "register within 12 months" and "register with 12 – 24 months." Both Discriminant Analysis and Logistic Regressions were run, but the models were unable to predict any projects as being a "yes" based on current data. In the end, a modified approach was used. The probabilities that each project would be a "yes" were computed out of a Logistic Regression, and those probabilities were then used to compute an Expected Value of each project. This provided us with a "ballpark" figure for GBCI to use to set its budget for the next year.

Technical Summary

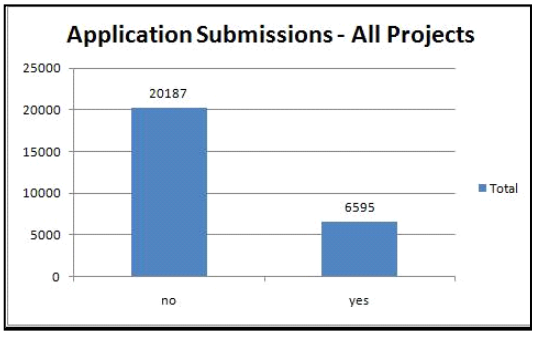
The data available for this analysis came from GBCI's SAP database, and was provided by GBCI as two separate exports with data as far back as 2000. The first file is a comprehensive spreadsheet with nearly 30,000 records and 22 columns. This file represents all projects that have at least registered. The second file is an export of all application transactions. These are the projects where GBCI has collected the application fee. This contains the date the fee was paid and is critical for the analysis. In order to analyze registration and application data together, these spreadsheets were merged using Spotfire through a unique identifier, 'Project ID.'

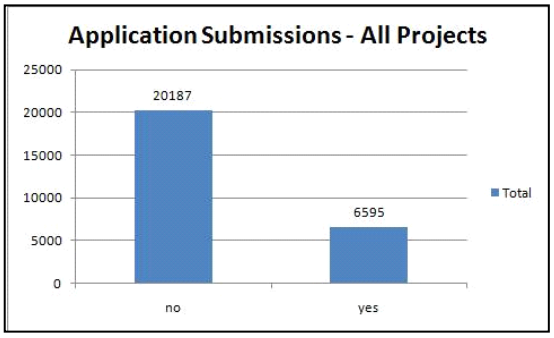
GBCI's data is complete for all certification applications, but is not without challenges. The most obvious is the fact that the variable to be predicted (Application Date) is present for a very small percentage of the total records (of the nearly 30,000 registered projects, only about 6,000 have applied for certification). To do any comprehensive analysis, the remaining (registered, but not yet applied) records must be considered. Additionally, the data is quite complex. For instance, there are five different types of LEED rating system (to measure five different types of buildings), each with a unique number of versions (for example LEED-NC has versions 2.0, 2.1, 2.2, and 2009, while LEED-Schools only has version 2007 and 2009).

The data did contain some outliers, particularly for project size. Some projects had entered '1' for building square footage, while others had entered '999999999'. This could be due to the fact that the project size was not known to the customer at the time of registration. Fortunately, this represented a small number of records and they were removed.

To begin the analysis we created some new variables. One was a variable that simply displayed whether or not a project had actually applied. Another included the length of time between registration date and the end of the certification version (systems became more stringent over time). Most importantly we captured the length of time between initial registration and application date for those projects that did apply. The summary statistics for that measure are provided in Appendix C.

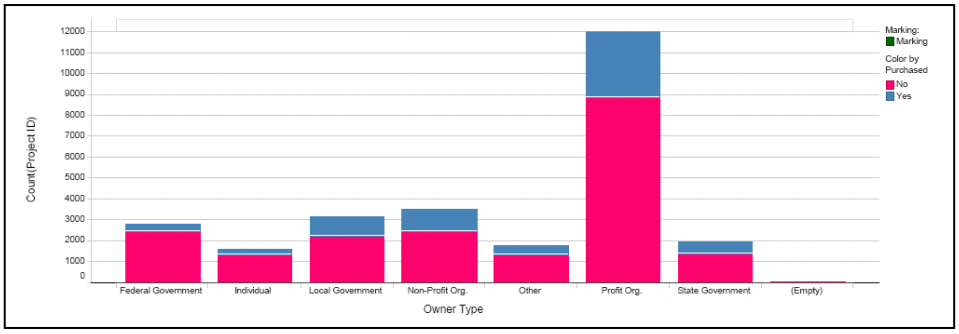
We then determined a few baseline statistics. First, the overall number of projects that historically moved from registration to application was determined, followed by summary statistics of how long on average it took. The total number of projects that moved from registration to application was 6,595, or 24.62% of all projects within the data set and can be seen in Appendix E – Application Rate. Of these 6,595, 33% submitted their application within 1 year of registration, 35% within 2 years, and a little less than 15% within 3 years. This can be seen in Appendix F. In all, roughly 83% of all projects that submitted an application did so within 3 years. Intervals of one, two and three years were chosen because GBCI does their budgeting on a yearly basis.

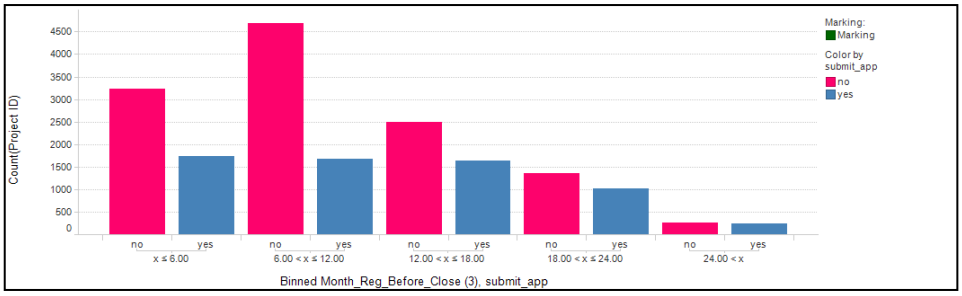


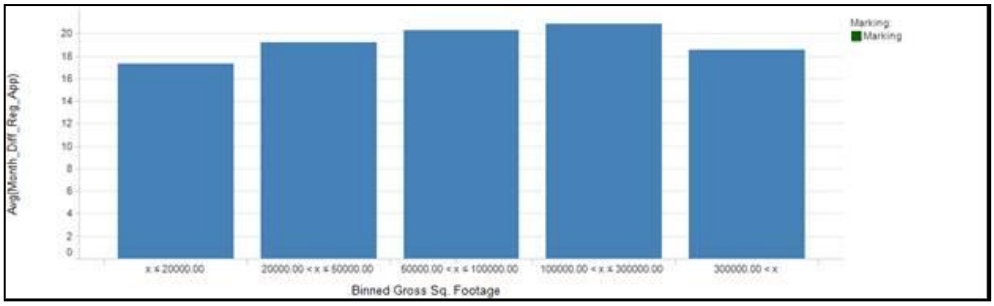


With these baselines determined, the nearly 30,000 records of LEED Certification were then explored to determine if there were any patterns that existed between variables such as Owner Type, Project Size and Months Before Registration Closes and whether or not a project actually applied. Variable types that seemed to have some variation were Owner Type, Gross Square Footage, and Months Before Registration Closes. To make the analysis more manageable, the Gross Square Footage column data was binned into 5 categories, and the Month Before Close was also binned.

Owner type showed a clear pattern in regards to whether or not a project submitted an application and how fast it took. Private organizations moved to the application stage at a higher rate than other groups, and did so faster than average. Not surprisingly, government projects, especially U.S. Federal Government, moved slower than average and at lower rates.







Additionally, we were able to see that the closer a registration came to the closing of a certification registration window the less likely the customer was to proceed to application. Dean used his domain knowledge to theorize that a lot of people register in the final days to be grandfathered in under older systems as the newer ones remove loopholes and become more stringent.

Despite the patterns that emerged, there was no "silver bullet" factor that allowed us to compute a straightforward model that would provide GBCI with an easy way to predict how many and which projects would move to the application stage within the next year. We initially tried a few models on the data. The most seriously considered model that wasn't used was a logistic regression for a numerical Y. We had attempted to predict the number of months between registration and application for GBCI. However, given the budgeting cycles there was an extra step required and our model's performance was not very compelling.

However, we felt that utilizing a Classification method for prediction would allow us to come up with a probability that each project would submit an application. The opportunity to use classification, and its associated probabilities, made a lot more sense for this application. There are a finite number of classes and the probabilities are useful for the business. While not perfect, a simple probability could then be used to assist in the budgeting process by allowing GBCI to compute an Expected Value of each project.

For the first step, we removed the data from 2009, as our first output variable would be "submit an application with 12 months." Since not all projects registered in 2009 had yet been "in process" for a full 12 months, their "no" responses might skew the data. With the 2009 data removed, we created dummy variables for owner type and binned square footage. A logistic regression was run with .5 as the cut-off value using a training and validation. Not surprisingly, the Logistic Regression classified all records as being a "No" within 12 months. However, our goal was not to predict whether these projects would be a "yes", but to get the probability, however low. To achieve a higher degree of accuracy, insignificant predictors were dropped and the model was re-run. The final predictors used in the model were Profit_Org, all of the square footage binned categories except for Greater than 300,000, and the Months before registration closes for between 6 to 9 and 9 to 12 months.

Once the probabilities were computed, we then looked at all registrations from 2008. Each record was assigned a probability based on the regression equation, and that probability was then multiplied by the fees each project would submit if it did indeed move to the next step. An expected value was then computed based on this data. The expected value based on a weighted average of fees (member/non-member) totaled around $6 million.

The above method could then be re-computed for the 12-24 month time period (excluding 2008 registrations), and the 24-36 month time period (excluding 2007 – 2009 registrations). In sum, the resource manager at GBCI could apply these methods to all outstanding projects at the end of year to help provide a picture of how much revenue would be generated, thus assisting with the budgeting process.

GBCI could also alter the cutoff value that we used in the classification if there is greater fund of over or underestimating budget. It did not appear that there was a substantial difference. We would recommend that GBCI use this model as input to its budgeting process. It could be compared against the 2009 budget to determine its accuracy there. Additionally, it could be tweaked in the future as more information comes in.

Appendix A – Leed Process

Appendix B – Data Dictionary

Variable	Type	Description
Confidential	Categorical	Yes/No response whether project wishes to remain confidential
Project ID	Unique String	Database assigned unique identifier
Project Name	Unique String	User-entered name
City	Categorical	City
State	Categorical	State
Country	Categorical	Country
LEED System Name	Categorical	Family of LEED
Version	Categorical	Version of LEED family
Next_Key_Event	Date	Last day to register under this version
Registration Date	Date	date project was registered
Certified?	Categorical	yes/no - is project certified
Certification Date	Date	Date project was certified
Certification Level	Categorical	Certification level achieved- certified, silver, gold, or platinum
Points Achieved	Continuous	out of a possible ~80, varies by version
Project Type	Categorical	Selected at registration - office, school, museum, etc
Gross Sq. Footage	Continuous	Project size
Owner Type	Categorical	profit, non-profit, gov't, etc
Owner Occupant Type	Categorical	profit, non-profit, gov't, etc
purchase type	Categorical	Review package (one phase or two)
Purchase Date	Continuous	date applied
Month_Diff_Reg_App	**Continuous**	**difference between registration and application**
Month_Reg_Before_Close	Continuous	difference between registration and close

Appendix C – Selected Summary Statistics of Months Between Registration & Application

Month_Diff_Reg_App	
Mean	19.0841
Standard Error	0.17184
Median	15.06452
Mode	9.645161
Standard Deviation	13.95504
Sample Variance	194.7432
Kurtosis	2.046966
Skewness	1.408769
Range	95.96667
Minimum	0
Maximum	95.96667
Sum	125859.6
Count	6595

Within 12 Months: Registration to Application

Count of Project ID	Column Labels		
Row Labels	no	YES	Grand Total
Federal Government	90.83%	9.17%	100.00%
Individual	91.12%	8.88%	100.00%
Local Government	91.91%	8.09%	100.00%
Non-Profit Org.	90.80%	9.20%	100.00%
Other	90.80%	9.20%	100.00%
Profit Org.	84.20%	15.80%	100.00%
State Government	92.92%	7.08%	100.00%
Grand Total	**87.99%**	**12.01%**	**100.00%**

Count of Project ID	Column Labels		
Row Labels	no	YES	Grand Total
x ≤ 20000.00	84.26%	15.74%	100.00%
20000.00 < x ≤ 50000.00	87.05%	12.95%	100.00%
50000.00 < x ≤ 100000.00	88.95%	11.05%	100.00%
100000.00 < x ≤ 300000.00	90.46%	9.54%	100.00%
300000.00 < x	90.64%	9.36%	100.00%
Grand Total	87.99%	12.01%	100.00%

Appendix D – Model Performance

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-2.89194608	0.07866113	0	*
Owner Type_Profit Org.	0.853589	0.06486281	0	2.34805894
Binned Gross Sq. Footage_20000.00 < x ≤ 50000.00	0.5533976	0.09024195	0	1.73915184
Binned Gross Sq. Footage_50000.00 < x ≤ 100000.00	0.38694692	0.09553575	0.00005116	1.47247839
Binned Gross Sq. Footage_x ≤ 20000.00	0.79020566	0.0784625	0	2.20384955
Binned Month_Reg_Before_Close (2)_6.00 < x ≤ 9.00	0.07650944	0.07935903	0.33499962	1.07951236
Binned Month_Reg_Before_Close (2)_9.00 < x ≤ 12.00	0.31825486	0.08451875	0.00016622	1.37472653

Residual df	9992
Residual Dev.	7210.158691
% Success in training data	12.33123312
# Iterations used	8
Multiple R-squared	0.03461841

Training Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	YES	no
YES	0	1233
no	0	8766

Error Report			
Class	# Cases	# Errors	% Error
YES	1233	1233	100.00
no	8766	0	0.00
Overall	9999	1233	12.33

Validation Data scoring - Summary Report

Cut off Prob.Val. for Success (Updatable)	0.5

Classification Confusion Matrix		
	Predicted Class	
Actual Class	YES	no
YES	0	969
no	0	7368

Error Report			
Class	# Cases	# Errors	% Error
YES	969	969	100.00
no	7368	0	0.00
Overall	8337	969	11.62

Appendix E – Sample Probabilities & Expected Values Per Project

Logit	Odds	Prob	Expected Value
-1.18078	0.30704	0.234913	557.9176579
-1.18078	0.30704	0.234913	557.9176579
-1.18078	0.30704	0.234913	557.9176579
-1.18078	0.30704	0.234913	557.9176579
-1.18078	0.30704	0.234913	557.9176579
-1.18078	0.30704	0.234913	557.9176579
-1.18078	0.30704	0.234913	557.9176579
-2.02202	0.132388	0.116911	277.6627832
-2.22901	0.107634	0.097175	230.7908728
-1.91651	0.14712	0.128251	304.59695
-1.91651	0.14712	0.128251	304.59695
-1.91651	0.14712	0.128251	304.59695
-1.91651	0.14712	0.128251	304.59695
-1.91651	0.14712	0.128251	304.59695
-1.91651	0.14712	0.128251	304.59695
-1.18078	0.30704	0.234913	557.9176579
-2.22901	0.107634	0.097175	230.7908728
-1.18078	0.30704	0.234913	557.9176579
-1.18078	0.30704	0.234913	557.9176579
-1.97095	0.139325	0.122287	290.4318315