# It’s not easy building green:

## predicting lead time for LEED certifications

May 10, 2010

GREEN BUILDING™
CERTIFICATION INSTITUTE

## Project Background

The Green Building Certification Institute (GBCI):

Independent, third party organization committed to promoting the design, development, and implementation of processes used to increase and measure green building performance and green building practice.

Provides two main services to its customers:

1) bestows onto individuals LEED Green Associate and LEED AP credentials

2) project certification, which involves administering the LEED certification program according to validated international standards of the LEED Rating System.

# Focus on the Certification Process :


•LEED certification involves three steps: Registration, Application and Certification.

•Certification process includes two revenue generation points, Registration and Application Review.

•Organizations seeking LEED certification pay an upfront registration fee as well as a second, larger fee for application review.

## Issue:

Inability to predict whether or not an registrant will file an application for review within 12 months of registration (GBCI's accounting cycle).

## Goal:

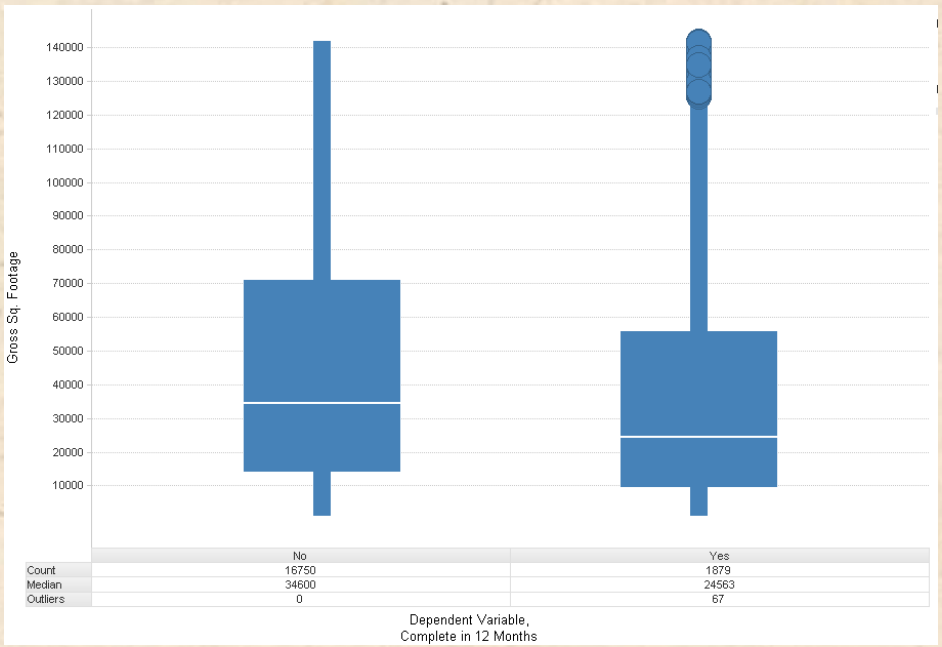
Evaluate various classification methods to tackle issue at hand.

An accurate prediction would allow the organization to perform more reliable cash flow forecasting, which will inform hiring and expansion plans.

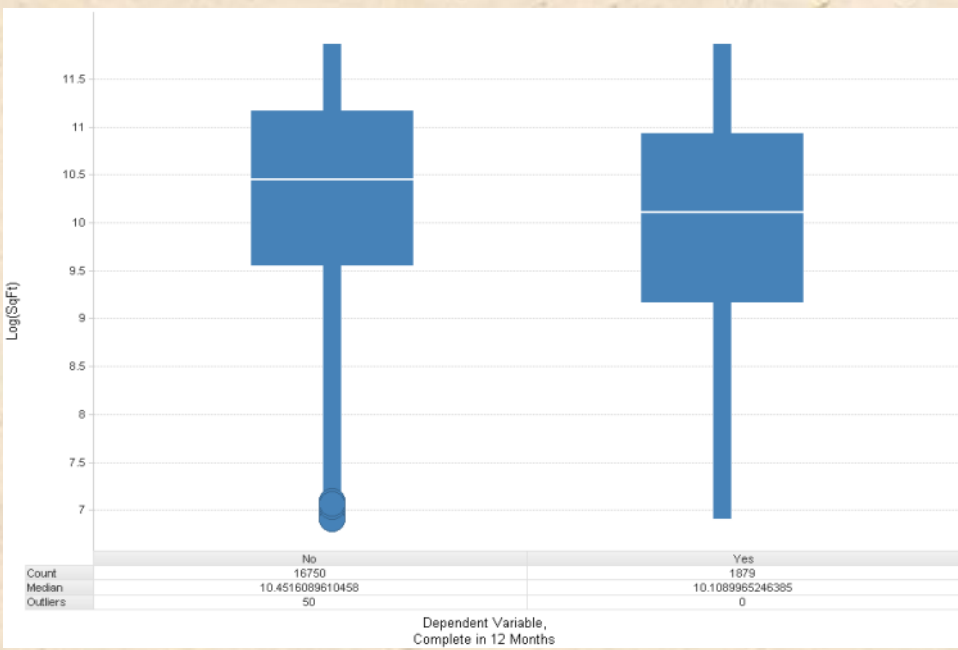
# A Look at the Sample Data

| Complete in 12 Months | Region | Proj_ Commercial | Proj_ MilitaryBase | ProjType | GrossSqFt | Owner Type | Month_Reg Before_Close |
|---|---|---|---|---|---|---|---|
| No | West | YES | NO | Commercial Office | 142604 | Profit Org. | 0.03 |
| Yes | South | YES | NO | Commercial Office | 160000 | Profit Org. | 0.03 |
| Yes | International | NO | NO | Campus | 230286 | Profit Org. | 10.77 |
| No | Midwest | YES | NO | Commercial Office | 20000 | State Government | 17.35 |
| No | South | NO | YES | Military Base | 37950 | Federal Government | 0.83 |
| No | West | NO | YES | Military Base | 9364 | Federal Government | 0.87 |

# Initial Exploration

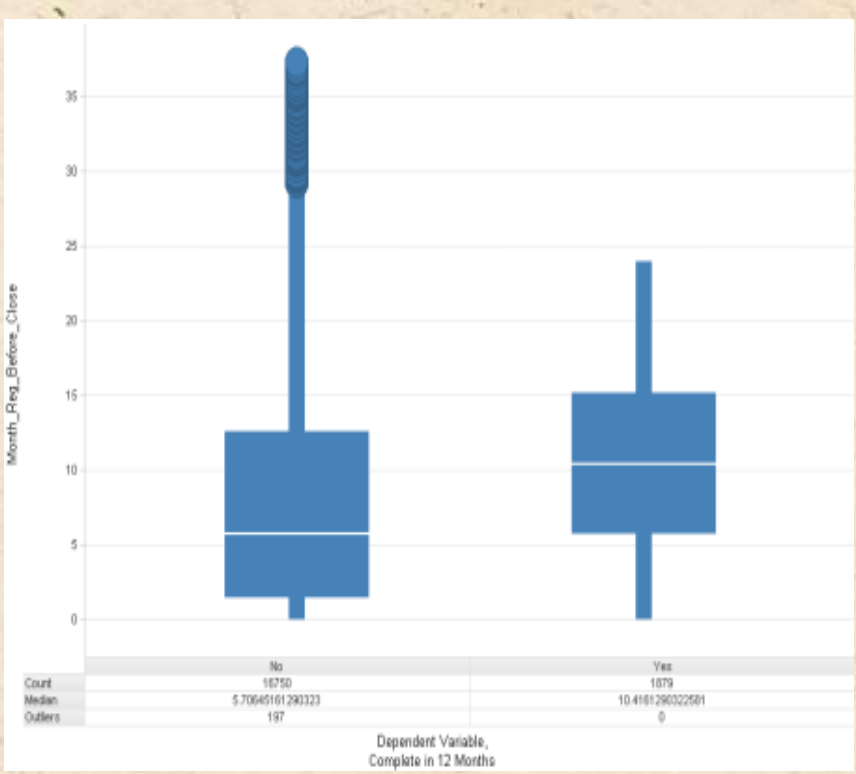



Square Feet



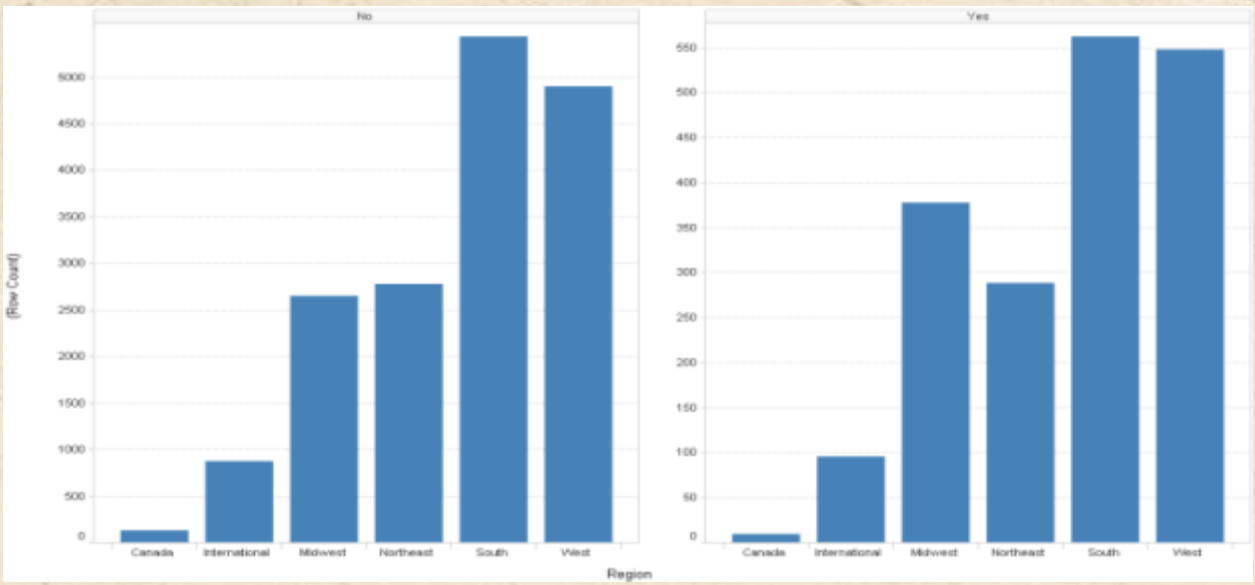

Log(Square Feet)

# Initial Exploration

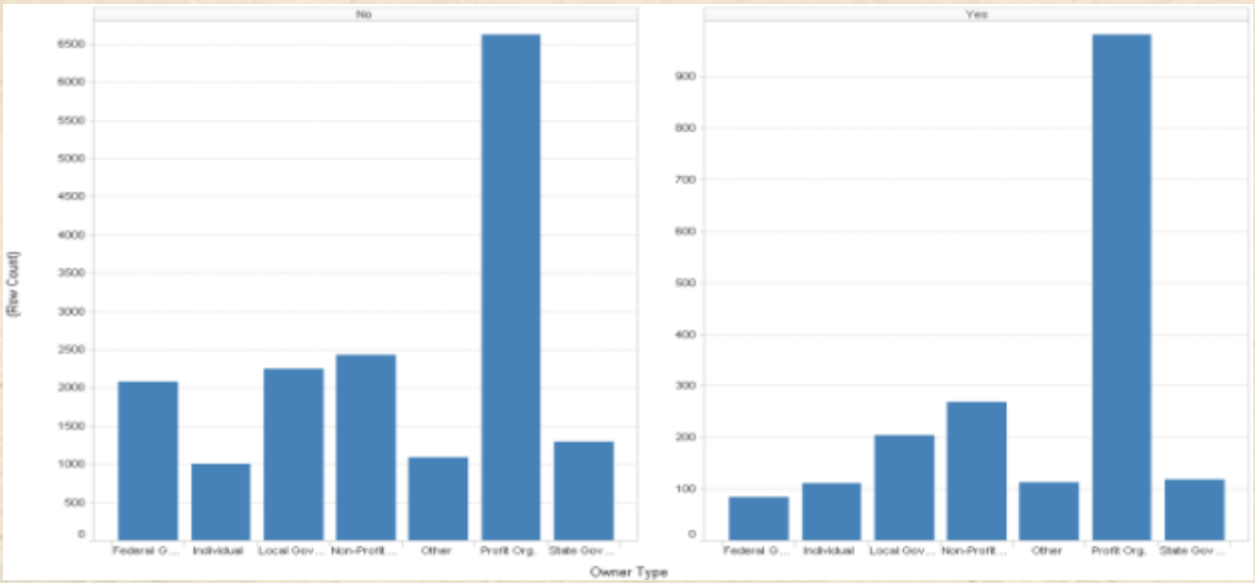



Time Until Close

# Initial Exploration




Regions




Owner Type

*The Models:*

Naïve Bayes, KNN,
CART & Logistic Regression

# Naïve Bayes Models

Reason for Use: would theoretically be able to integrate the given predictors into the naïve rule to generate accurate classifications

Variables Utilized: Confidentiality, Region, Owner Type, Occupant Type, and Square Footage; Square footage values binned, as the model does not allow for continuous variables

- 10 separate, evenly spaced bins– 400,000 sq.ft. in breadth
- 4 bins, separated by visualization cut off

Analysis based upon: Conditional probabilities, Classification Confusion Matrices, Overall error %

Role of Oversampling

Result: variables provided did not adequately capture the behavior of the observations provided.

# Naïve Bayes

**Conditional probabilities**

| | Classes--> | | | |
|---|---|---|---|---|
| Input Variables | Yes | | No | |
| | Value | Prob | Value | Prob |
| Confidential | No | 0.775862069 | No | 0.709764475 |
| | Yes | 0.224137931 | Yes | 0.290235525 |
| Region | Canada | 0.00862069 | Canada | 0.008832188 |
| | International | 0.062807882 | International | 0.084028459 |
| | Midwest | 0.204433498 | Midwest | 0.146344455 |
| | Northeast | 0.139162562 | Northeast | 0.166952895 |
| | South | 0.285714286 | South | 0.309371933 |
| | West | 0.299261084 | West | 0.284470069 |
| BinData2 | 1 | 0.208128079 | 1 | 0.122791953 |
| | 2 | 0.168719212 | 2 | 0.161678116 |
| | 3 | 0.213054187 | 3 | 0.186825319 |
| | 4 | 0.410098522 | 4 | 0.528704612 |
| Owner Occupant Type | al Government | 0.054187192 | Federal Government | 0.110279686 |
| | Individual | 0.051724138 | Individual | 0.056673209 |
| | al Government | 0.088206897 | Local Government | 0.10954367 |
| | ed Occupancy | 0.195812808 | Mixed Occupancy | 0.227060844 |
| | Non-Profit Org. | 0.156403941 | Non-Profit Org. | 0.134200196 |
| | Profit Org. | 0.406403941 | Profit Org. | 0.294774289 |
| | e Government | 0.049261084 | State Government | 0.067468106 |

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | Yes | No |
| Yes | 0 | 812 |
| No | 0 | 8152 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| Yes | 812 | 812 | 100.00 |
| No | 8152 | 0 | 0.00 |
| Overall | 8964 | 812 | 9.06 |

## Conditional probabilities

| | Classes--> | | | |
|---|---|---|---|---|
| Input Variables | Yes | | No | |
| | Value | Prob | Value | Prob |
| Reg_June_YES | 0 | 0.886086249 | 0 | 0.741253051 |
| | 1 | 0.113913751 | 1 | 0.258746949 |
| Proj_Commercial_YES | 0 | 0.397884459 | 0 | 0.608624898 |
| | 1 | 0.602115541 | 1 | 0.391375102 |
| SqFtBinned | 1 | 0.213181448 | 1 | 0.130187144 |
| | 2 | 0.193653377 | 2 | 0.181448332 |
| | 3 | 0.18633035 | 3 | 0.167615948 |
| | 4 | 0.406834825 | 4 | 0.520748576 |

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | Yes | No |
| Yes | 763 | 466 |
| No | 452 | 777 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | %Error |
| Yes | 1229 | 466 | 37.92 |
| No | 1229 | 452 | 36.78 |
| Overall | 2458 | 918 | 37.35 |

# Naïve Bayes, Oversampling

# KNN

- Various ‘k’ values examined
- Model complicated by fact that all but two variables were categorical
- Oversampling Employed

KNN Results: variables provided did not adequately capture the behavior of the observations provided.

# KNN Models

### Validation error log for different k

| Value of k | % Error Training | % Error Validation | |
|---|---|---|---|
| 1 | 0.03 | 16.79 | |
| 2 | 8.15 | 10.11 | |
| 3 | 8.24 | 11.26 | |
| 4 | 9.06 | 9.54 | |
| 5 | 9.00 | 9.86 | |
| 6 | 9.20 | 9.53 | |
| 7 | 9.24 | 9.63 | |
| 8 | 9.27 | 9.35 | <--- Best k |
| 9 | 9.27 | 9.37 | |
| 10 | 9.29 | 9.35 | |

### Validation error log for different k

| Value of k | % Error Training | % Error Validation | |
|---|---|---|---|
| 1 | 0.04 | 46.18 | |
| 2 | 22.99 | 65.14 | |
| 3 | 23.23 | 45.48 | |
| 4 | 29.66 | 59.54 | |
| 5 | 29.50 | 45.42 | <--- Best k |
| 6 | 31.98 | 57.40 | |
| 7 | 32.71 | 45.52 | |
| 8 | 33.97 | 56.22 | |
| 9 | 35.44 | 46.13 | |
| 10 | 36.90 | 52.44 | |

### Test Data scoring - Summary Report (for k=8)

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

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | No | Yes |
| No | 6177 | 2 |
| Yes | 640 | 0 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| No | 6179 | 2 | 0.03 |
| Yes | 640 | 640 | 100.00 |
| Overall | 6819 | 642 | 9.41 |

### Test Data scoring - Summary Report (for k=5)

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

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | Yes | No |
| Yes | 306 | 308 |
| No | 2594 | 3346 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| Yes | 614 | 308 | 50.16 |
| No | 5940 | 2594 | 43.67 |
| Overall | 6554 | 2902 | 44.28 |

# CART

**Reason for Use**: Given the large number of potential dummy variables, looking for a modeling process that might clue us in to the most important dummies in a category, or the appropriate bins to use for continuous variables.

**Result:**

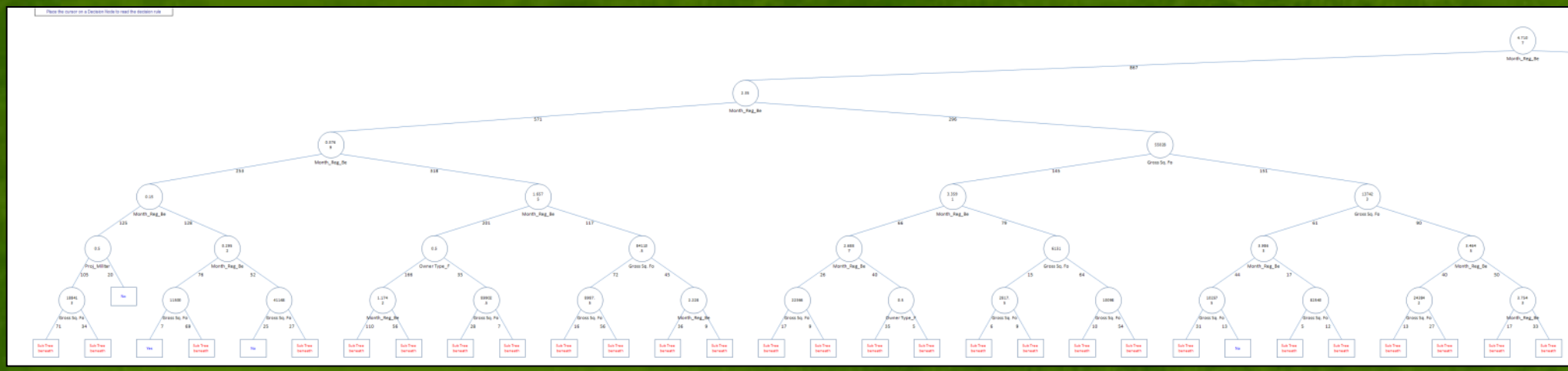
Both XLMiner and Spotfire Miner returned, as a tree, a single box with the word “No”. After eliminating the forced pruning of the tree, “Yes” nodes were generated, but at level 7 within the tree.

“Months_Diff_Reg_Close” supporting the initial impressions that it would be an important factor in distinguishing between “Yes” records and “No” records.  It is not a clear enough signal, however, to create a practical tree.

# Minimum Error Tree (Naïve Case)

No

## Partial screen shot of un-pruned tree



# Logistic Regression Models

Reason for Use: predict a categorical response variable also in hope to understand the odds of the success class

Modeling: Various models were tried before conclusion

- Original Model with the data at hand
- Modified data: condensing the project type to- Commercial, Military
- Modified SqFt to LOG(SqFt)

Role of Oversampling:

- Predictive Accuracy of success class increased with the tradeoff of decreasing overall accuracy.

Result:

- Variables do have a small p value suggesting their significance
- Odds close to one suggest the signal is not strong enough for a definite conclusive model

# Logistic Regression Models

**Test Data scoring - Summary Report**

| Cut off Prob.Val. for Success (Updatable) | 0.25 |
|---|---|

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | Yes | No |
| Yes | 4 | 636 |
| No | 26 | 6188 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | % Error |
| Yes | 640 | 636 | 99.38 |
| No | 6214 | 26 | 0.42 |
| Overall | 6854 | 662 | 9.66 |

Modified Data- Project Type

**Test Data scoring - Summary Report**

| Cut off Prob.Val. for Success (Updatable) | 0.25 |
|---|---|

| Classification Confusion Matrix | | |
|---|---|---|
| | Predicted Class | |
| Actual Class | Yes | No |
| Yes | 597 | 17 |
| No | 5076 | 900 |

| Error Report | | | |
|---|---|---|---|
| Class | # Cases | # Errors | %Error |
| Yes | 614 | 17 | 2.77 |
| No | 5976 | 5076 | 84.94 |
| Overall | 6590 | 5093 | 77.28 |

Over Sampled Data

# Model Summaries

| Model | Best Error % Test Data | Notes |
|---|---|---|
| Naive-Bayes | 9.34 | Standard Partition (all classified as "No") |
| CART | 9.33 | Standard Partition (all classified as "No") |
| Logistic Regression | 9.66 | Standard Partition (all classified as "No") |
| KNN | 9.37 | Standard Partition, k=8 (10 total) |

| Model | Best Error % Test Data | Notes |
|---|---|---|
| Naive-Bayes | 37.35 | Oversampled Partition (40% 'Success') |
| CART | 9.32 | Oversampled Partition (40% 'Success') |
| Logistic Regression | 77.28 | Oversampled Partition (40% 'Success') |
| KNN | 44.29 | Oversampled Partition (40% 'Success'), |

Initial Findings:

Lower square footage, more time from registration to the end of an application system, and residing in the Midwest all increase the probability of a “Yes” outcome.

However…

**We advise GBCI not to use these variables for predicting** time between registration and application. Classification models were unable to demonstrate that these variables have sufficient predictive ability.

The reason: likely that the drivers behind application lead time are not yet reflected in the available data.

Final Recommendation: Modify the registration process to include two questions, whose responses might serve as more discriminating predictive variables.

1. What is the completion status of your project? (Idea, Plans Drafted, Construction in Progress, Completed)

2. When do you expect to submit the completed application forms? (0-3, 6-12, 12-24, 25+ months)

# Questions?