



House Investment

Profit in 30 days or less

BUDT733
5/10/07

**John Bradshaw
Laura Gonzalez
Richard Liao
Bharat Menghani
Michael Movesian**

Executive Summary

The goal of our project was to try to use data available to realtors to help a potential investor predict which houses will sell in under 30 days in the towns of Germantown and Gaithersburg. This information would be very useful to a real-estate investor who wanted to generate a profit quickly and avoid paying any mortgage payments by selling under 30 days.

After organizing and cleaning our data, we performed a visual analysis using boxplots, histograms, and scatterplots in an attempt to gain some insights into the housing information. This helped us choose what to include in our statistical models.

We tried applying a Logistic Regression, Classification Tree, and Discriminant Analysis to the data. Our analysis showed that the Classification Tree only produced the non-success class. The Logistic Regression's output had minimal lift. Our optimal Discriminant Analysis provided some lift for the 14 most likely houses to sell less than 30 days. The model predicted that these 14 houses would sell less than 30 days, but in reality only 5 of those houses sold that quickly. Even so, an investor could use this top ranked list to help narrow down the list of homes to consider. Since an investor will only be investing in a few houses at a given time, it is critical to choose the right property and prevent costly mistakes.

Thus none of our models were accurate enough to be used as a predictive tool by themselves. This led us to believe that there must be some other factors, which we did not consider, that would cause a house to sell in less than 30 days. We speculated on such possible factors that could be examined in a future study to determine if they could be used to more accurately predict whether a house would be sold in under 30 days.

We could gain some further insights by speaking with experts in the field who possess the domain knowledge. In addition, additional predictors that are valued by domain experts could be included to improve the accuracy of models.

Technical Summary

Our goal is to predict whether a house will sell within 30 days. This information is useful to an investor who wishes to flip a home quickly in order to make a large profit. If the investor could select a house with certain characteristics she would be able to use the model to predict whether the house would sell in under 30 days.

Methodology

We received our data from the Metropolitan Regional Information System (MRIS) database of public listing houses. We selected data from January 1st, 2006 through December 31st, 2006 since this was the year that the housing market slowed considerably. Next, we looked at our dataset for errors, cleaned the data, and removed outliers. We decided to eliminate houses built prior to 1955 since these homes were likely to have been built with asbestos. Also, we will not be flipping old vintage homes by remodeling them, since that would cause them to lose value. We deleted listings with Lot Sq Ft area greater than or equal to 20,000 because the land might be worth more than the house, and lots larger than 20,000 square feet are outliers.

For the purposes of modeling, our success class is "Sold in under 30 days". We chose a time period of 30 days since mortgage payments would begin after that. Based on the Naïve Rule, our success class is 19%, so we want a model that can predict better than that.

Visualization

We first used visualization tools to look for relevant predictors in our data. By looking at boxplots of predictors vs. DaysUntilSale, we found that Bedrooms, Levels, Age and List Month showed separation so they might be useful predictors in our model (Exhibit 1).

We binned age based on some domain knowledge. A histogram showed that homes in the range of 20-25 and 30-35 years seem to sell a little faster than other aged homes based on the percentage (Exhibit 2). Also, we found that homes older than 40 years sell the fastest.

A histogram showed that homes listed between September and December sold in under 30 days less often than homes listed during the rest of the year (Exhibit 3). Thus, we decided to create a dummy variable (List_Fall) showing whether a home was listed in September through December or not. We also saw that a lower percentage of homes in Germantown sold in under 30 days during the months September through December than homes in Gaithersburg (Exhibit 4). So we created an interaction term between city and List_Fall.

Models

Classification Tree

We ran a Classification Tree on all the predictors. This resulted in a full tree that had all non-success class as the leaves so we cannot use this tree to predict. The predictors at the top were Style_Colonial, Bedrooms and List_Fall. The predictors at the top of the tree gave us some insight as to what predictors might be significant in the models. Indeed we found that Style_Colonial and List_Fall showed up again in the models.

Logistic Regression (LR)

We ran a Logistic Regression, eliminated variables with high p-values and ended up with a model with two predictors: Style_Colonial, and BathsF*BathsH. The model still had the maximum error rate of 100% for the success class (IE sold under 30 days) and an r-squared of 0.017. This model had no predictive abilities and was of no use to an investor since it had no lift (Exhibit 6). Thus we decided not to use this model.

Discriminant Analysis (DA)

Next, we standardized the data and ran the Discriminant Analysis which allowed us to compare classification scores and compare variable relevance. This led to the four most relevant predictors of: Style_Colonial, Style_Rambler, List_Fall, and List_Fall * City. (Exhibit 7) We noticed that Style_Colonial and List_Fall also showed up Classification Tree and Logistic Regression. So it seems clear that these variables seem fairly important.

Although the model has a high error rate, the model has enough lift to be valuable for prediction. The model predicted 14 properties had 67% chance of selling in under 30 days. Of these, 5 actually sold in under 30 days. The model had a 35.7% accuracy rate for the top 14 performers (Exhibit 8), which is better than the 19% in the dataset as a whole.

Analysis

In general our analyses show some interesting insights. In the summer when the market is at its peak there is no difference the rate at which houses sell between the two towns. Also homes that list in the fall are less likely to close in under 30 days than homes listed in other times of year. Homes that listed in the fall are more likely to sell in under 30 days in Gaithersburg than Germantown. A higher percentage of Rambler Style homes sell in under 30 days than Colonial homes.

Conclusions

The DA model is valuable and performs better than the Naïve Rule. Predictors in the model are easy and inexpensive to obtain from the MRIS database. However, it does not predict accurately on its own and we recommend it be used as a funnel to help narrow down the number of houses that an investor would consider buying in order to resell in less than 30 days. In addition Domain knowledge should be used to enhance the accuracy of the selected homes.

In order to improve the model we feel that we need to capture additional predictors suggested by the seasoned investors who possess the domain knowledge. They could make a more accurate prediction of investment properties that will sell in under 30 days.

The model could have been improved with the addition of the following predictors

1. School data
2. How was the house advertised
3. Crime data
4. Agent ID
5. Prime Mortgage Rate
6. Garage (1 or 2)
7. Curb Appeal (rating system)

Exhibit 1 - BoxPlots

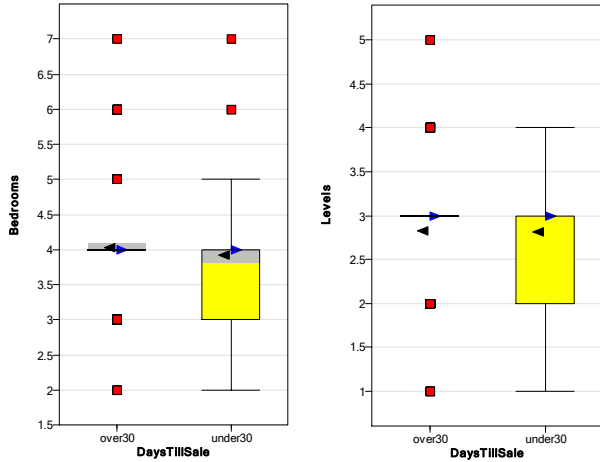
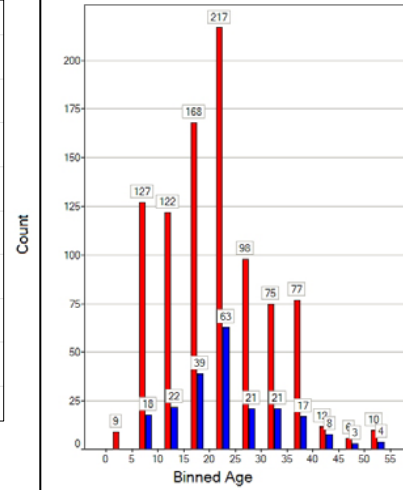


Exhibit 2 – Visualization with ListPrice



Age	Over 30	Under 30	Ratio
0-5	9	0	0.0%
5-10	127	18	12.4%
10-15	122	22	15.3%
15-20	168	39	18.8%
20-25	217	63	22.5%
25-30	98	21	17.6%
30-35	75	21	21.9%
35-40	77	17	18.1%
40-45	12	8	40.0%
45-50	6	3	33.3%
50+	10	4	28.6%

The height of a bar represents the number of records.
 Color by DaysTillSale:
 over30 (red), under30 (blue)
 The x-axis is binned into 12 bins.
 The labels show the height of each bar.

Exhibit 3 – ListMonth

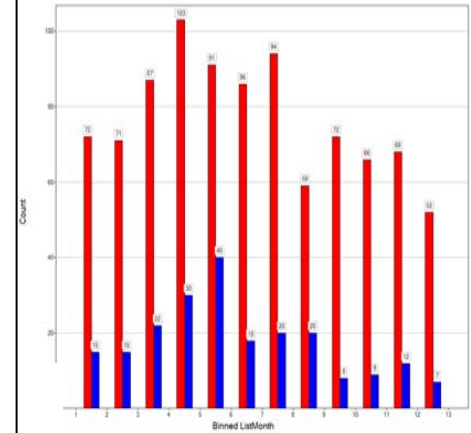


Exhibit 4 – ListMonth Trellis by City

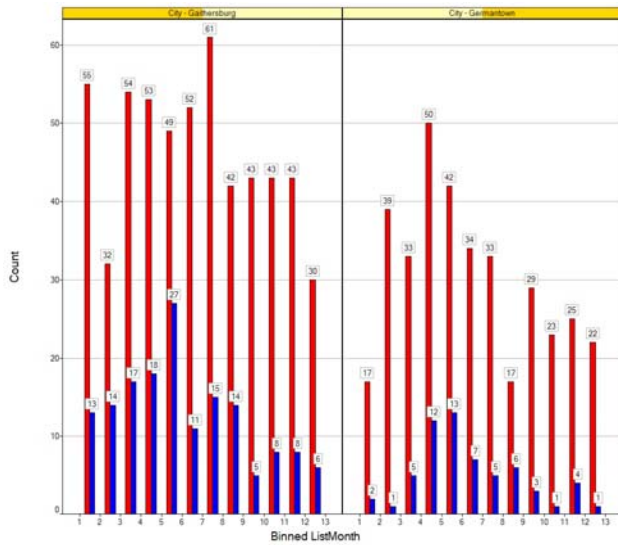


Exhibit 5 – Visualization with ListMonth

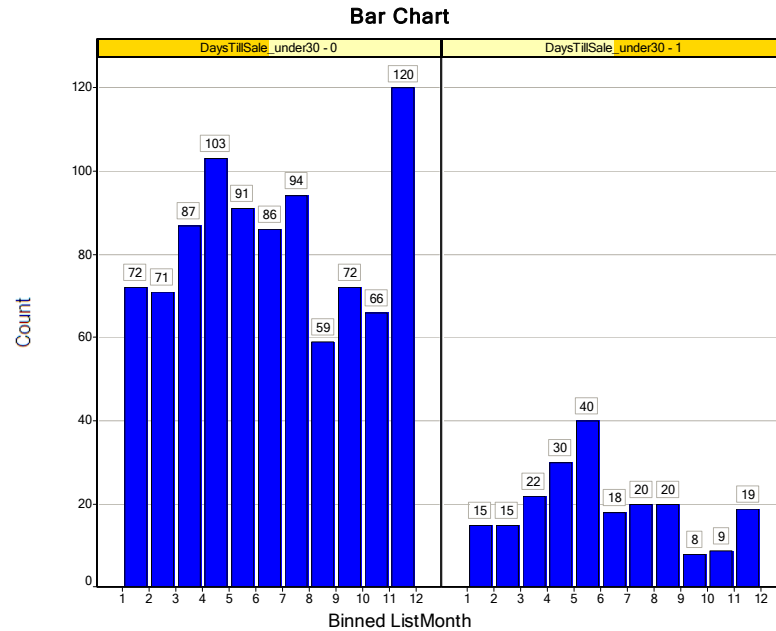


Exhibit 6 – Optimal Logistic Regression

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-1.33921933	0.2263152	0	*
Style_Colonial	-0.86422586	0.27808383	0.00188493	0.42137763
BathsF*BathsH	0.2303381	0.11205607	0.03982484	1.25902557

Test Data scoring - Summary Report

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	0	46
0	1	181

Error Report			
Class	# Cases	# Errors	% Error
1	46	46	100.00
0	182	1	0.55
Overall	228	47	20.61

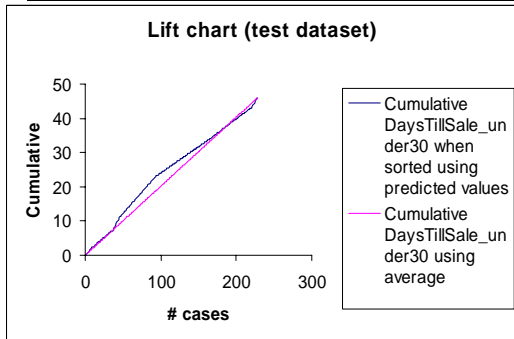


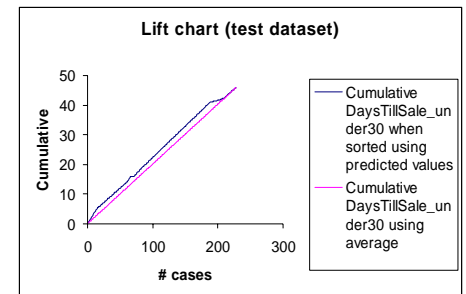
Exhibit 7 – Optimal Discriminant Analysis

Classification Function

Variables	Classification Function	
	1	0
Constant	2.31985235	2.69493675
Style_Colonial	4.18903112	4.64750719
Style_Rambler	4.79105377	4.44759274
List_Fall	0.94362676	1.40053689
List_Fall * City	-0.3429364	0.37485161

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	16	30
0	50	132

Error Report			
Class	# Cases	# Errors	% Error
1	46	30	65.22
0	182	50	27.47
Overall	228	80	35.09



Variable	Definitions
ListPrice	The list price.
Style_Colonial	Homes that are colonials.
Style_Rambler	Homes that are Ramblers.
Bedrooms	Number of bedrooms
BathsFull	Number of full bathrooms
BathsHalf	Number of half bathrooms
BathsF*BathsH	BathsFull multiplied by BathsHalf
Levels	Number of levels
Fireplace_y/n	Whether the home has any fireplaces
Basement Y/N	Whether there is a basement
Lot Sqft	The area of the property
Age	The age of the home in years
DaysTillSale	Sold under 30 days or not
List_Fall	Whether home listed in September – December

Exhibit 8 – Discriminant Analysis Test Score Report, Top Ranked Properties

Cut off Prob.Val. for Success (Updatable)	0.5	(Updating the value here will NOT update value in summary report)
---	------------	---

Row Id.	Predicted Class	Actual Class	Prob. for 1 (success)	Style_Colonial	Style_Rambler	List_Fall	List_Fall * City
59	1	0	0.672	0	1	0	0
70	1	0	0.672	0	1	0	0
120	1	0	0.672	0	1	0	0
196	1	0	0.672	0	1	0	0
329	1	0	0.672	0	1	0	0
542	1	0	0.672	0	1	0	0
550	1	0	0.672	0	1	0	0
654	1	0	0.672	0	1	0	0
890	1	0	0.672	0	1	0	0
953	1	1	0.672	0	1	0	0
979	1	1	0.672	0	1	0	0
1006	1	1	0.672	0	1	0	0
1121	1	1	0.672	0	1	0	0
1135	1	1	0.672	0	1	0	0