Determining Factors of a Quick Sale in Arlington's Condo Market

Team 2:
Darik Gossa
Roger Moncarz
Jeff Robinson
Chris Frohlich
James Haas

Executive Summary

The real estate market for condominiums in Northern Virginia has been extremely volatile over the past several years. At the height of the real estate boom, properties put on the market were sold within a matter of days, if not hours. As the market has cooled, the number of days that a property remains on the market has risen. With that in mind, our goal was to create an explanatory model that would help us determine which factors lead to a quick sale of a condominium in Arlington, so realtors could market condos more effectively. For purposes of this analysis, a ‘quick sale’ is defined as a property going under contract within two weeks after it is listed.

Our primary data source was the Metropolitan Regional Information Systems (MRIS) database, which serves as the primary listing service for the real estate industry in the Mid-Atlantic region. Working closely with a real estate agent specializing in Northern Virginia, we extracted data from the MRIS database on all condominiums sold in Arlington between October 25, 2006 and October 24, 2007. Our final dataset included more than 1,600 records that contained pertinent information such as list price, close price, and any subsidies offered by the seller. In addition, each record contained key attributes of a condominium, such as the total square footage, number of bedrooms and bathrooms, and the age of the building. We supplemented information on each condo to include data about the surrounding community. For example, we collected data on the level of crime within a radius of the property, as well as its proximity to parks, retail establishments, and the Metro. We also examined current local foreclosure data to establish whether the spate of recent foreclosures might have deterred the quick sale of condominiums, but ultimately determined that trying to link these to the time of condo sales proved prohibitively difficult given our time constraints.

After collecting this data and narrowing our variable set through exploratory analyses, we ran classification trees and logistic regression models to determine the factors that most influenced a ‘quick sale’. Ultimately we selected a 6-predictor logistic regression model for its clarity in this explanatory task. As it turns out, this model demonstrated a poor fit to the data with little improvement over the Naïve model, high error rates, and low sensitivity in identifying “quick sale” condos. While this result was disappointing to us, all 6 the predictors in our model did display statistical significance. From this, we can conclude that the condos that were more likely to be quick sales:

- Were located farther from convenience shopping
- Were located farther from Metro Stations
- Were older rather than newer properties
- Were surrounded by fewer parks w/in .25 Miles
- Displayed less of a drop between List Price and the total amount paid by the buyer
- Commanded higher close prices per square foot

If we were to perform this analysis again, we would change the following two factors: 1) vary the amount of time for a ‘quick sale’, and 2) include more macro-economic factors into the model.

In defining two weeks as a ‘quick sale’, we were heavily influenced by our domain knowledge and summary statistics of our data set. For example, our domain knowledge led us to pick two weeks as a ‘quick sale’, since that was how quickly apartments sold during the last year. (Roughly 30% of the condos in our data set were classified as ‘quick sales’.) However, the real estate market over the past several years has been abnormally strong due to macroeconomic influences. A ‘quick sale’ during the past year was only two weeks, where as a ‘quick sale’ in previous years would normally be classified as less than one month. If we varied the ‘quick sale’ definition, the other property features might have had a greater impact.

A second approach might involve collecting data on economic indicators for real estate such as: existing-home sales; new-home sales; housing starts; foreclosures; and employment levels among others. To the extent that future researchers can find local or regional values for these indicators and link them to the period when a given condo was on the market, they are more likely to be successful in finding a model that provides a better fit and lower error rates.

Technical Summary

Variable Selection and Analysis

In addition to property attributes, we also considered community factors that might affect a 'quick sale' of condo properties in Arlington. Specifically, we collected and measured both the distance to and also the count of local retail businesses (grocery and convenience stores), parks, crime, nightlife (bars and restaurants), and Metro stops in the vicinity of each property, using a process called Geocoding. Geocoding provided global coordinates for each location, thereby allowing us to calculate distances between each point, and the number of establishments within a given radius. After Geocoding each property and its community attributes, each property had approximately 80 variables, which could potentially affect a 'quick sale'. Therefore, before we could create a model to profile a 'quick sale', we first needed to eliminate the meaningless variables from the analysis.

To do this, we divided the variables into similar data groups, and then subjected each data group to the variable reduction process summarized in **Exhibit 1.** First, we eliminated variables that did not make sense using domain knowledge. Next, we separated the results into two classes ('quick sale' and non-'quick sale') and visually explored each variable using box plots, scatter plots, and summary statistics to eliminate variables that did not exhibit separation between the two classes. Then, we eliminated variables that contained similar information by viewing a correlation matrix of the remaining variables in each group. Finally, we eliminated variables whose accuracy was questionable. In end, this process yielded the following 9 variables: Count_AllParks_pt25M, Dist_ClosestMetro_M, Dist_ClosestConv_M, CondoAge, PriceMovement, PricePerSqFt, TotSqFt, Count_RestBarHotel_pt25M, and Count_CrimeTotal_pt5M. (Definitions of each variable can be found in **Exhibit 1**.)

Modeling Process

In determining the most important factors influencing a 'quick sale', we experimented with several different types of classifiers including discriminant analysis, classification trees, and logistic regression. Ultimately we selected logistic regression because of the clarity of the results for our explanatory task.

First, we ran a logistic regression model on all of our variables remaining after the data reduction step described above. After running the initial model, we then tried several combinations of variables using both P-values and domain knowledge to guide our selections. This allowed us to test many possible explanations for quick condo sale before arriving at our ultimate model.

The ultimate logistic regression model we selected contained 6 predictors, all of which were statistically significant at a 95% confidence interval. The details of this model are displayed in **Exhibit 2**, and it contained the following variables: Count_AllParks_pt25M, Dist_ClosestMetro_M, DistClosestConv_M, CondoAge, PriceMovement and PricePerSqFt. Of the variables that remained from our exploratory analysis, Count_RestBarHotel_pt25M, Count_CrimeTotal_pt5M, and TotSqFt were excluded from the ultimate model.

Model Performance

Although we were able to identify several variables for our final model that were statistically significant in explaining quick sales of condominiums in Arlington, overall our model produced a rather poor fit to the training data. This is reflected in our low Multiple-R squared value of .07297, indicating there is only an approximate 7% improvement over the naïve model in determining 'quick sales'. Other shortcomings of the model are apparent when examining its low sensitivity (15.26%) and high error rate (84.77%) in correctly identifying the success class of 'quick sales'. While the specificity of our model was high (95.91%), and its overall error rate (28.31%) somewhat lower than that for the 'quick sale' condos, we must conclude that our model does not effectively describe the factors that affect a 'quick sale'.

Findings on Individual Explanatory Variables

Despite its poor performance overall, our model did identify 6 factors that explain quick sales of condominiums in Arlington with statistical significance at the 95% confidence interval. Each of these had a different effect on the odds of a condo being sold within 14 days and thus considered “quick”. Among the variables included in our model, one that provided a surprising finding was the distance to the closest metro (Dist_ClosestMetro_M). The output for this variable indicated that condos further from Metro stations were more likely to sell quickly than condos near stations. This was indicated by the positive coefficient (0.26891452) for the distance to metro variable. Given this coefficient, for a condo that is a quarter of a mile farther from a metro station, the odds of it selling quickly increased by approximately 7% as reflected by the corresponding odds factor of 1.069539979

A second variable that produced a surprising result was the distance to the closest convenience store (Dist_ClosestConv_M). When collecting our data, we expected that condos with convenience stores close by would be more likely to sell quickly. However, the coefficient for this variable was 0.98084885, indicating that for a quarter-mile increase in the distance from its nearest convenience store, its odds of a ‘quick sale’ increased about 28% as reflected by the corresponding odds factor of 1.27789246.

Another surprise emerged from the model output when we examined the age of the condos (CondoAge). While we expected that newer condos would sell quickly, increases in the age of a condo also increased the likelihood that it would sell quickly. The regression coefficient of 0.01129689 indicated that 10 year increase in the age of a condo increased its odds of selling ‘quickly’ by ~12% as shown by the corresponding odds factor value of 1.119597113.

Another interesting finding was that the price variables did not have a large effect on how quickly the condo sold. To measure the difference between asking price and the final price paid, we created the PriceMovement variable by subtracting the close price and any seller subsidy from the original list price. This variable corresponded with our expectations that over-priced condos would not sell quickly since the seller needed to concede more off the asking price in order to sell. However, the coefficient for this variable was only -0.00003258. Given this, if a buyer ultimately paid $1000 below list price for the condo, the odds of a quick sale decreased by approximately 3.2% with an odds factor of 0.967945011.

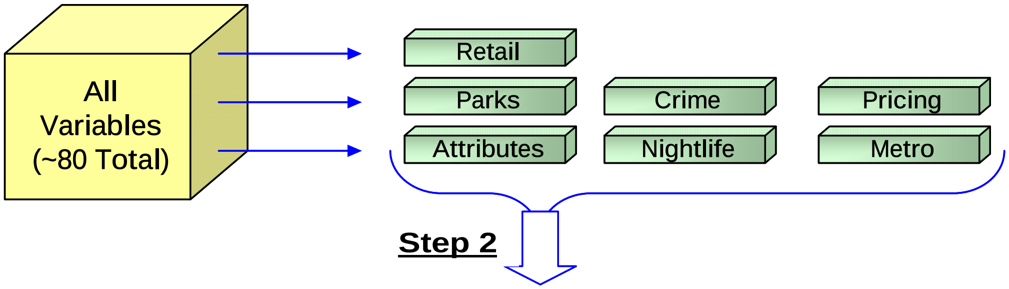
The second variable used to measure price was PricePerSqFt, which was created by dividing the total square footage by the effective closing price (closing price – seller subsidy). Since price per square foot is used as a proxy for condo quality, we expected that condos with a higher value for price per square foot would be more attractive to buyers, and thus sell more quickly. Similar to the price movement variable however, the impact of changes to the price per square foot on the odds of a quick sale were somewhat limited; the coefficient for this variable was only 0.00639065. To put this in perspective, the average total square foot measure for our sample of condos was 1040 sq ft. Increasing the closing price per square foot by $1 would therefore increase the closing price of the condo by about $1000 on average. This $1 increase in the closing price/sq ft would increase the odds of the condo selling quickly by only 0.6% as demonstrated by an odds factor of 1.00641108.

The final variable included in our model was the number of parks found within a quarter mile of the property (Count_AllParks_pt25M). This actually had a negative effect on the odds of a condo being sold quickly. This was reflected in the coefficient value of -0.14204456, and the corresponding odds factor value of 0.86758262. Thus, for every additional park found within this quarter-mile radius, the odds of the condo being sold quickly declined by approximately 13%. These findings ran contrary to our expectation that a high density of parks close to the property would make the condo more attractive to buyers, and thus increase its chances of being sold quickly. Given this, we feel there may be other location-based factors at work here despite the predictor’s significance.

Exhibit 1:

Explanatory Variable Analysis and Selection Process

Step 1



Step 2

Explanatory Variable Reduction Process

Domain Knowledge Test	Does this variable make sense or is it explainable?
Visual Exploratory Analysis	Does a box plot or scatter plot reveal separation btw the classes?
Compare AVE and STDEV	Eliminate variables that have Similar AVEs and High STDEVs
Correlation Test	Are variables within group highly correlated?
Accuracy Test	Is the variable created with accurate and reliable data?

Result

Remaining Variables for Profiling Model

Variable	DataType	FieldType	Description
Count_AllParks_pt25M	Integer	Community Features	Number of parks within .25 miles of the condo
Dist_ClosestMetro_M	Decimal	Metro	Dist (miles) to closest metro station (any line)
Count_RestBarHotel_pt25M	Integer	Nightlife	Number of restaurants/bars/hotels within .25 miles of condo
Dist_ClosestConv_M	Decimal	Retail	Dist (miles) to closest convenience store
Count_CrimeTotal_pt5M	Integer	Crime	Number of crimes (all types) within .5 miles
PriceMovement	Money	Pricing	Equal to ClosePrice - Subsidy - ListPrice. If positive, condo sold above list, if negative, condo sold below list.
TotSqFt	Integer	Condo Attributes	Square footage size of the condo
CondoAge	Integer	Condo Attributes	Age of the condo (years). 2007-YrBuilt
PricePerSqFt	Money	Condo Attributes	The Real Close Price (ClosePrice-Subsidy) amount divided by Total Sq Ft

Step 1 – Separate Variables into "like category" groups
Step 2 – Subject each group of variables to the Reduction process
Result – Best remaining variables available to subject to the Profiling model

Examples of Visual Exploratory Analysis

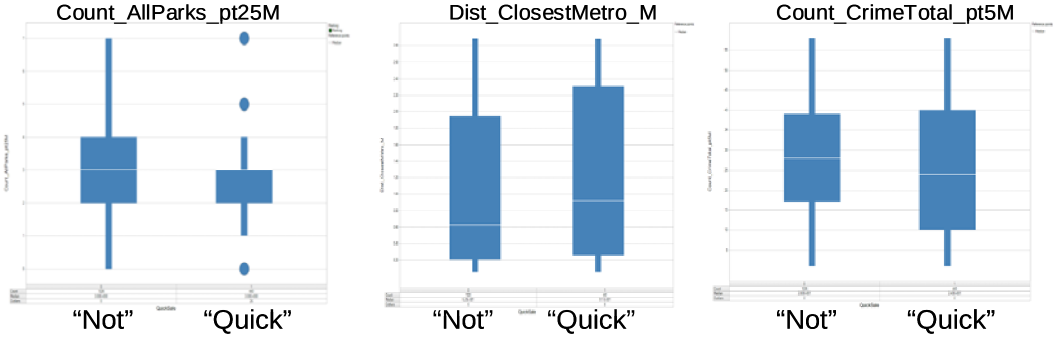


Exhibit 2

Inputs

Data	
Training data used for building the model	['Final Models.xls']'Data'!$A$2:$S$1467
# Records in the training data	1466

Variables						
# Input Variables	6					
Input variables	Count_AllParks_pt25M	Dist_ClosestMetro_M	Dist_ClosestConv_M	Age	L-C-S	C-S/Tsq
Output variable	QuickSale					
Constant term present	Yes					

Parameters/Options	
# Iterations	50
Marquardt overshoot factor	1
Initial cutoff probability value	0.5
Confidence Level %	95

Output options chosen
Summary report of scoring on training data

Prior class probabilities

According to relative occurrences in training data

Class	Prob.	
1	0.300136426	<-- Success Class
0	0.699863574	

The Regression Model

Input variables	Coefficient	Std. Error	p-value	Odds
Constant term	-3.70876455	0.47157034	0	*
Count_AllParks_pt25M	-0.14204456	0.05453635	0.00919857	0.86758262
Dist_ClosestMetro_M	0.26891452	0.08197037	0.00103575	1.30854332
Dist_ClosestConv_M	0.98084885	0.45526034	0.03120263	2.66671896
CondoAge	0.01129689	0.00324709	0.00050314	1.01136088
PriceMovement	-0.00003258	0.00000497	0	0.9999674
PricePerSqFt	0.00639065	0.00087736	0	1.00641108

Residual df	1459
Residual Dev.	1660.669312
% Success in training data	30.01364256
# Iterations used	9
Multiple R-squared	0.07297314

Training Data scoring - Summary Report

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	67	373
0	42	984

Error Report			
Class	# Cases	# Errors	% Error
1	440	373	84.77
0	1026	42	4.09
Overall	1466	415	28.31