UNIVERSITY OF MARYLAND

BUDT 733: DATA MINING FOR BUSINESS

PREDICTING ZILLOW'S ZESTIMATE ACCURACY

PREPARED BY TEAM # 6

Sirisha Kilambi
Pramod Mane
Beena Nair
Mehul Shah
Tabrez S Shaikh

TABLE OF CONTENTS

Executive Summary

Zillow.com is a popular online real estate website which gets around 8 million hits a month, assisting users in every stage of the home ownership process—buying, selling, remodeling and financing. A few years ago, Zillow created a breakthrough when it introduced its free instant online appraisal known as "Zestimate" and became the go-to place for customers to obtain this information instantly before going through the complicated appraisal process involving banks, brokers, financial professionals etc.

The goal of our project was to assess whether the Zestimate values for homes located in 17 select cities of Northern Virginia, such as Herndon, Reston, Arlington etc. were accurate within +/-5% of the final market value. This helps 2 primary stakeholders - buyers and sellers. Buyers can determine if Zillow is accurate about the price of the neighborhood and bid the right price for a prospective home. It also helps sellers set their home price right which is the key to selling their home successfully in the market.

The data was collected from various websites like Zillow, Redfin, Schooldigger and google maps. Zillow and Redfin were used to obtain data pertaining to the home, such as location, home type, bedrooms, bathrooms, square feet, total rooms, lot size, date of last sale etc. Schooldigger gave information about the ranking of the primary, middle and high schools for each house listing in the data set. The proximity to metro (in miles) was computed from Google maps.

Our analysis used a dataset of 1416 records to generate the binary response variable "ZestimatesCorrect" that assumes the values "True" or "False" depending upon whether the Zestimate was within +/-5% of the final market value. A variety of classification techniques were applied and the KNN model gave the best results. According to this model, the user needs to input the following 5 variables - Squarefoot, lot size (in sqft), total number of rooms, number of bathroooms and middle school ranking. Based on these input variables, our model will predict whether the Zestimate provided by Zillow is accurate or not. The model is not only parsimonious but also the data points that it needs are easily available to the end user. Also, this model can be extended to any other city in Northern Virginia as the final model is independent of the city in which the house is listed. These advantages override the fact that the model may be computationally a little more expensive than other models. The other models turned out to be a little more complicated and difficult to use from an end user standpoint than the selected model. The overall accuracy of this model is around 70%.

This tool can be marketed to real estate agents/brokers who are the main contact points for sellers and buyers in the market. For a false outcome for a particular listing, the agent can use his local expertise and adjust the intended price of the house. This could help spur home sales and boost the local economy by aligning the interests of buyers and sellers. This tool could also be directly consumed by individual buyers or sellers enabling them to achieve their objectives.

By investing more resources in terms of time and money, information on other attributes like tile type used (Ceramic more expensive than linoleum), kitchen improvements, bonus-room, finished basement square footage etc can also be captured to improve the overall accuracy of the model. Also, given more time, the number of data points collected could be increased to improve the accuracy of data-driven models like KNN.

Technical Summary

Data Preparation for analysis

The data was collected for a time period between Jan 1, 2006 through September 2009 as this was inclusive of the boom and bust period in the housing market. The missing values of the predictor variables from the data sources were merged to get a more robust and complete data set. A snapshot of the data collected is given in Appendix A. A new variable ZestimateCorrect was defined and used as the predicted boolean response variable in the analysis. It was true if the Zestimate value at the time of sale of a house was within +/-5% range of the sale price. Based on the naïve rule the data set had an overall error rate of 33% on the full dataset of 4 years (3576 records). Our initial data exploration and output of certain models prompted us to get rid of records with missing data to bring the overall data set down to 1416 records. This also allowed us to get a higher percentage of records with correct zillow prediction (about 38%).

Data Exploration

Spotfire was used for data exploration and visualization of the multivariate data. A series of scatter plots and box plots were generated to explore the data. Due to the higher number of variables it was not very easy to speculate on which variables would have the best relationship. In the end we moved away from scatter plots and relied more heavily on box plots that allowed us to visually compare the variation of multiple variables across the true and false categories.

After the initial exploration, we found that missing values in records caused too much noise to see any kind of patterns. Since we had sufficient data, we could eliminate rows with lot of missing data. Replotting with this new dataset helped us discern better patterns. We also got rid of variables such as exterior_rev, which did not show any predictive power based on box plots. This reduced our data set from 3576 to 1416 records, which was still a sufficiently large data set.

The following variables seemed interesting based on the box plots because they showed difference in medians for the two categories. Bathroom_rev, Sqft_rev, log(sqft_rev), log (lot_size_rev), Age_of_house_at_sale, MiddleSchoolRank, TotalRooms_rev (Appendices : B, C and D)

We also ran principal component analysis on the data set (Appendix F) and found that the above variables had better weights than the non-interesting variables in the components that explained most variances. In some of our models we tried other seemingly insignificant variables also, but in the end those variables got eliminated based on p-values or position in the classification tree.

Choice of Metrics for Model Selection

Our goal was to predict if Zillow's estimate for house value was accurate (within +/-5 % of the price at which the house will likely sell). Given this task, we used the following criteria to choose the best possible model:

- Accuracy in light of misclassification costs. Our model could be wrong in two ways – it predicts that Zillow is correct when in reality it is not; or, the model predicts that Zillow is wrong and in reality it is correct. We took the perspective of various stakeholders including Sellers, Buyers and realtors. We concluded that the costs of misclassification were symmetric for our task. Therefore, instead of looking at just specificity or sensitivity, choosing the model with least overall error made sense. Also, we did not make lift charts or ranking part of evaluation because our problem does not require the top tier.
- Lower error rate on test data compared to the Naïve rule applied to the test data

- Parsimony of models with equivalent error rates.
- Availability and relevance of the model variables at the time of prediction:

Very early in our data exploration we found that variables such as "Age of Zillow" may be a good predictor of the zestimate accuracy but it did not really matter in the practical world because the variable would have the same value for all houses presented in a new test data set at any given point in time. While we sifted out such variables in data preparation, this criterion was retained for evaluating the models to ensure that we did not overlook the practicality of using the model in real life scenarios.

Models Execution and Results Interpretation

We partitioned the data into training (60%), validation (20%) and test (20%) datasets because some of the models we used (classification tree and KNN) used the validation set to optimize the initial model. Out of the models that were tried, the following three were interesting and had the lowest overall error rate.

Classification Tree:

This was our initial model (see Appendix E for detailed output) and we got slightly lower error rate of 34.28% compared to Naïve error of 35% on the test data set. This model also reassured us about the interesting variables chosen in the data exploration phase. The best pruned tree included lot_size and distance_from_metro. Based on our model performance criteria this model was acceptable because it had better accuracy than naïve rule on test data, was parsimonious, and had variables that were available and relevant at the time of prediction.

Logistic Regression:

This model (see Appendix E for detailed output) gave us even lower error rate of 32.86% compared to Naïve error of 35% on the test data set. The final model had BATHROOM_REV, LOG(SQFT), log(LOT_SIZE), TOTALROOMS_REV, Age_of_house_at_Sale, Binned_PrimarySchoolRank, Binned_MiddleSchoolRank and Binned_HighSchoolRank variables. Based on our model performance criteria this model was acceptable because it had better accuracy than naïve rule on test data, was parsimonious, and had variables that were available and relevant at the time of prediction.

K Nearest Neighbors:

Since KNN does not give insight into what variables are important, we used the knowledge from data exploration and prior models to select the input variables. The final model had BATHROOM_REV, SQFT_REV, LOT_SIZE, TOTALROOMS_REV, MiddleSchoolRank. This model (see Appendix E for detailed output) gave us even lower error rate of 31.91% compared to Naïve error of 35% on the test data set. Based on our model performance criteria this model was acceptable because it had better accuracy than naïve rule on test data, was parsimonious, and had variables that were available and relevant at the time of prediction.

Recommendation

All our models provide some improvements over the Naïve classification and have realistic parsimonious input variable set. However, we recommend the use of KNN model as by far it provides the best predictive accuracy. This model is data driven and can be used for solving the problem of predictions for houses in Northern Virginia. The practical implication will be that we will have to refresh the data periodically and also provide a way for users to check school rankings.

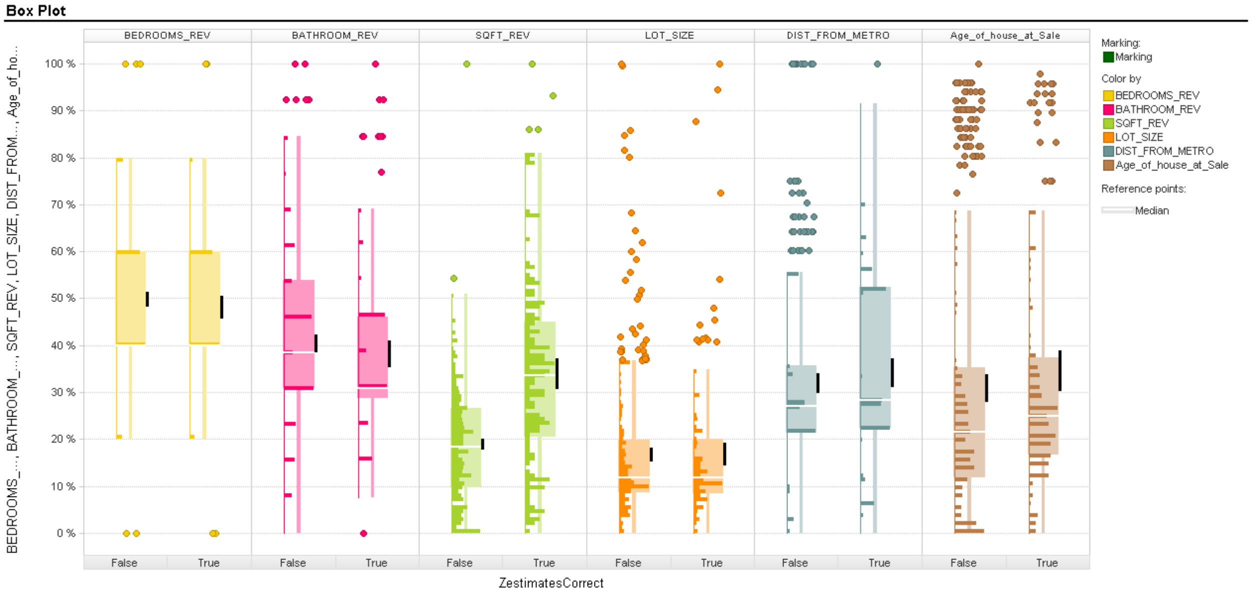
Appendices

Appendix A: Variables used for analysis.

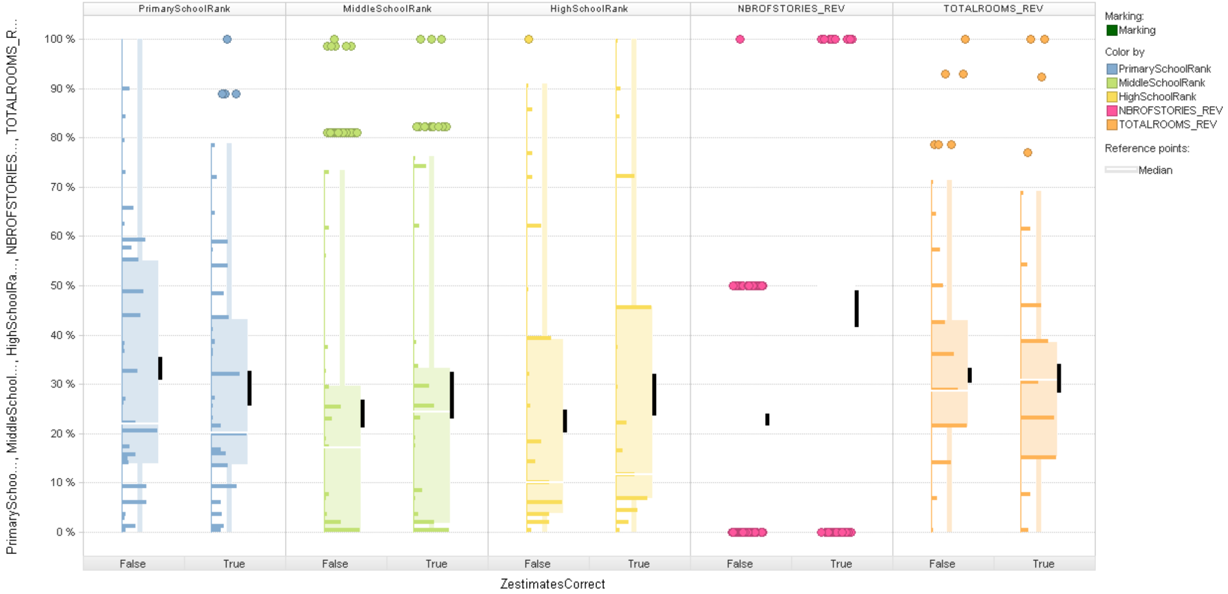
Variable	Definition	Variable	Definition
Home_Type	SingleFamily/Townhouse etc	***RoofType***	Roof Type
City	City	***Exterior material***	Siding,brick etc
Zip Code	Zip Code	***Season of sale***	Summer,Fall etc
Bedrooms	Number of Bedrooms	***Architecture***	Colonial, Ranch
Bathrooms	Number of Bathrooms	***Number_Of_Stories***	Number of stories
SQFT	Area in sqft	***Primary School Rank***	Primary School Rank
Lot Size	Lot size in sqft	***Middle School Rank***	Middle School Rank
Age of House	Age of the house	***High School Rank***	High School Rank
ZestimateCorrect	True or False	***Distance from metro***	Distance from metro
Basement	Finished/Unfinished	***Recession Period***	Yes/No
TotalRooms	Total number of rooms	***Parking Type***	Garage/Open

Excluded variables due to insignificance: Exterior material, Basement, Architecture, Roof Type, Parking type

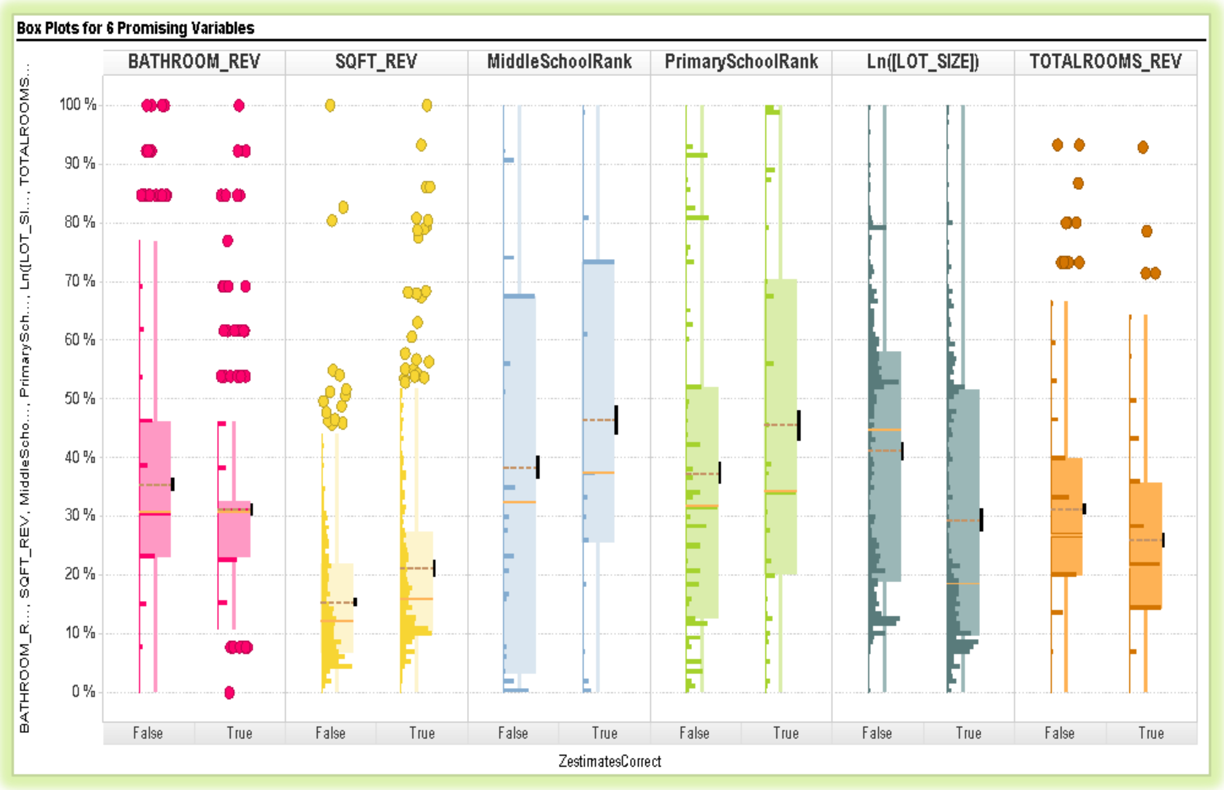
Appendix B : Box Plots of variables



Appendix C : Box Plots of variables



Appendix D: Final variables of interest



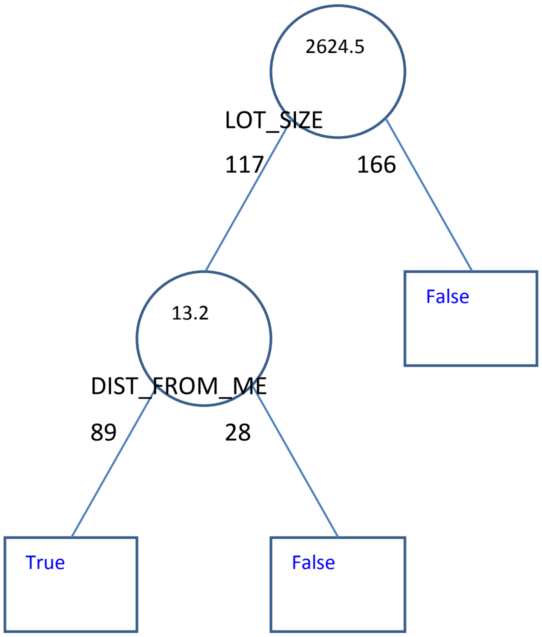
Appendix E: Model Outputs

Classification Tree

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	FALSE	TRUE
FALSE	143	41
TRUE	56	43

Error Report			
Class	# Cases	# Errors	% Error
FALSE	184	41	22.28
TRUE	99	56	56.57
Overall	283	97	34.28



K- Nearest Neighbors

Variables					
# Input Variables	5				
Input variables	BATHROOM_REV	SQFT_REV	LOT_SIZE	TOTALROOMS_REV	MiddleSchoolRank
Output variable	ZestimatesCorrect				

Parameters/Options	
# Nearest neighbors	20

Cut off Prob.Val. for Success (Updatable)	0.5	
Classification Confusion Matrix		
	Predicted Class	
Actual Class	FALSE	TRUE
FALSE	144	39
TRUE	51	48

Error Report			
Class	# Cases	# Errors	% Error
FALSE	0	0	Undefined
TRUE	0	0	Undefined
Overall	0	0	Undefined

Logistic Regression

Input variables	Coefficient	Std. Error	p-value	Odds	95% Confidence Interval	
Constant term	-4.65478611	2.11161542	0.0274982	0.009515948	0.008876227	0.01015567
BATHROOM_REV	0.38922957	0.12691256	0.00216283	1.47584331	1.15083241	1.89264178
LOG(SQFT)	0.2396526	0.35761046	0.50276226	1.27080762	0.63049442	2.5614059
log(LOT_SIZE)	0.38037464	0.12761253	0.00287591	1.46283257	1.13912296	1.87853193
TOTALROOMS_REV	-0.19049983	0.04766456	0.00006424	0.82654589	0.75282639	0.90748429
Age_of_house_at_Sale	0.01936915	0.01151659	0.09259845	1.01955795	0.99680215	1.04283321
Binned_PrimarySchoolRank	0.0735151	0.10634829	0.48939756	1.07628477	0.87378258	1.32571769
Binned_MiddleSchoolRank	-0.09299159	0.10067139	0.35563511	0.91120118	0.74803621	1.10995638
Binned_HighSchoolRank	0.04271848	0.12909032	0.74070543	1.04364407	0.81034607	1.34410834

Residual df	841
Residual Dev.	1069.3762
% Success in training data	61.647059
# Iterations used	11
Multiple R-squared	0.0551556

Cut off Prob.Val. for Success (Updatable)		0.5
Classification Confusion Matrix		
	Predicted Class	
Actual Class	FALSE	TRUE
FALSE	162	22
TRUE	71	28

Error Report			
Class	# Cases	# Errors	% Error
FALSE	184	22	11.96
TRUE	99	71	71.72
Overall	283	93	32.86

Appendix F : Principal Component Analysis Results

Variable	1	2	3	4	5	6	7	8	9	10	11	12
BEDROOMS_REV	0.2068	0.0196	0.1593	0.1412	0.0052	-0.1804	-0.7134	-0.4692	0.8757	-0.4789	-0.4350	0.4906
BATHROOM_REV	0.2001	0.1669	-0.1032	-0.0303	0.1228	0.2858	-0.3169	0.2257	-0.1167	-0.1253	1.5834	0.1837
SQFT_REV	0.2127	0.1597	0.0241	0.0334	0.0463	0.2260	0.1357	0.0717	0.0114	-0.1546	-0.4948	-1.7639
LOT_SIZE	0.1944	0.0076	0.2787	0.1569	0.0471	0.3033	0.7239	0.6361	0.3112	0.3582	-0.3498	0.8503
NBROFSTORIES_REV	0.0089	0.2659	-0.4365	-0.1938	-0.2363	-0.4937	0.5689	-0.0173	0.7735	0.0242	0.1307	0.0824
TOTALROOMS_REV	0.1719	0.0879	0.2295	0.1619	-0.2646	-0.7971	0.1314	-0.1980	-0.9032	0.1873	0.1012	0.1166
DIST_FROM_METRO	-0.0778	0.1652	0.3947	-0.2822	-0.7219	0.2973	-0.2571	-0.0765	0.2304	0.5754	0.1520	-0.1298
PrimarySchoolRank	-0.1434	0.2025	0.0389	0.3308	0.1407	-0.3623	-0.5886	0.9271	0.1966	0.3422	-0.1329	-0.2150
MiddleSchoolRank	-0.1529	0.2255	0.3313	0.0313	0.0008	0.0290	0.3306	0.0844	-0.0967	-1.4154	0.1161	0.1296
HighSchoolRank	-0.1103	0.2280	0.1508	0.3835	0.4045	0.1412	0.2800	-0.8097	0.1944	0.6638	0.3049	-0.0693
Age_of_house_at_Sale	-0.0296	-0.3391	0.2878	0.1156	-0.0452	-0.3674	0.2847	0.1048	0.6830	-0.0474	0.9038	-0.7507
salesvolume	0.0166	0.0652	0.2980	-0.6541	0.6254	-0.3112	-0.0512	0.0503	0.0469	0.2577	-0.0788	-0.0041

Variance	3.8646	2.3663	1.2212	1.1252	0.8045	0.6216	0.4594	0.4421	0.3449	0.2946	0.2462	0.2093
Variance%	32.2051	19.7189	10.1763	9.3768	6.7042	5.1799	3.8284	3.6843	2.8743	2.4554	2.0520	1.7444
Cum%	32.2051	51.9240	62.1003	71.4771	78.1813	83.3612	87.1897	90.8739	93.7482	96.2036	98.2556	100.0000
P-value	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0095	1.0000