PREDICTING ZILLOW.COM -ZESTIMATE'S ACCURACY

Group – 6

- Sirisha Kilambi
- Pramod Mane
- Beena Nair
- Mehul Shah
- Tabrez Shaikh

Agenda

- Introduction
- Data – Collection & Description
- Exploratory Analysis
- Modeling Methods
- Results
- Conclusion
- Questions

Introduction

Background

- Housing has been the root cause behind the current recession which is worst ever, next only to great Depression
- Recovery hinges on housing which in turn depends on buyers & sellers getting the right deal in right timeframe.

Core Idea

- "Zillow.com" is the real estate service launched in 2006
- Zillow calculates a Zestimate - home valuation as a starting point for anyone to see — for free — for most homes in the U.S
- Zillow indicates that for MD and VA they get only about 26% of predictions within the +/-5% range only.

Goal: To create a **Predictive** model that will provide the buyers and sellers with a tool to assess if zillow's estimate is reliable for the house they are trying to list or Buy.

Data Collection & Description

Collection

•Data collected, cleansed and merged from 4 sources – Zillow , Redfin, School Digger and Google Maps

•17 counties (29 Zip codes) in Northern VA

House sales data

Before Data Clean up: **3500+**

After Data Clean up: **1416**

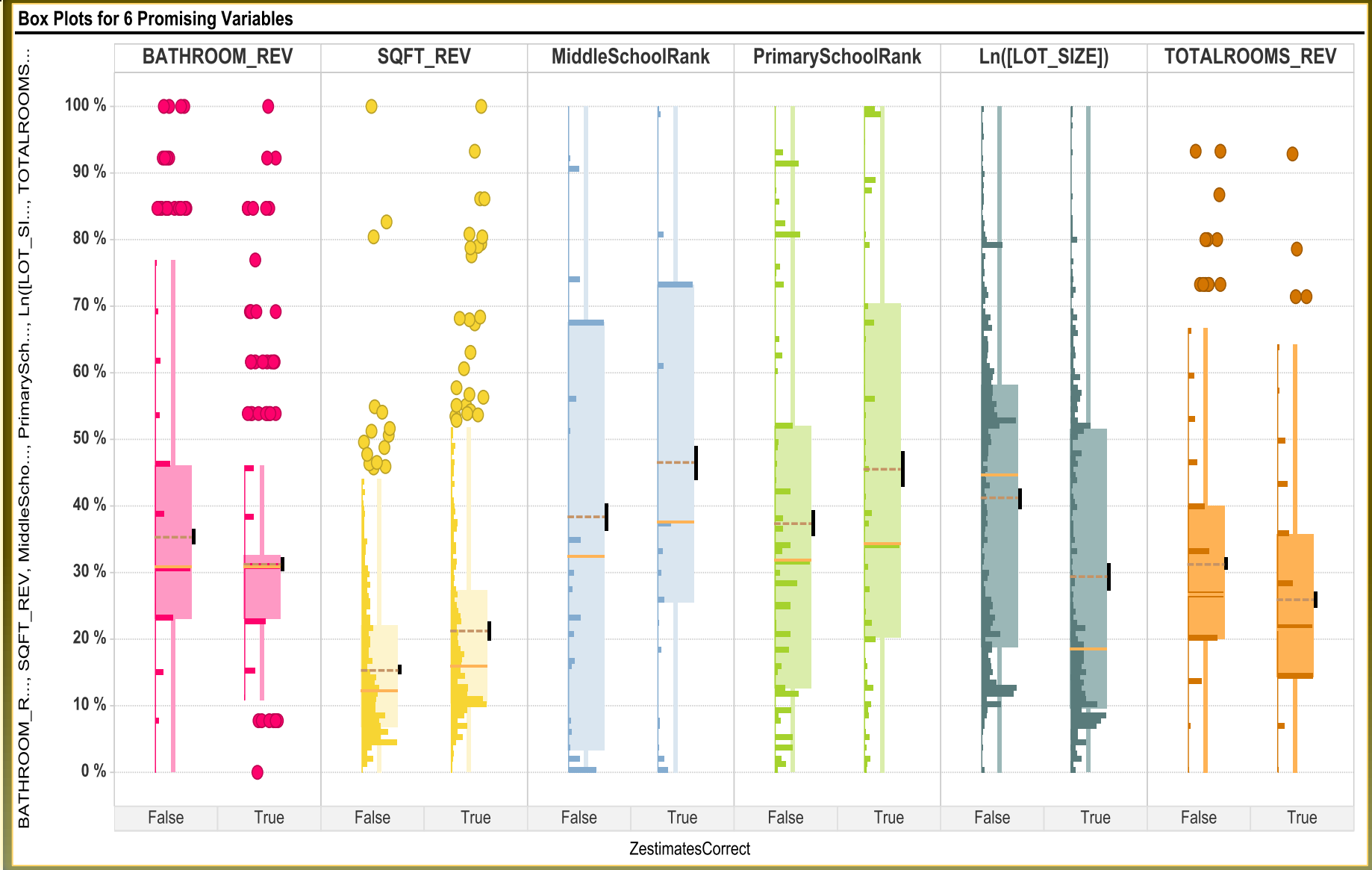
Y – ***Is Zestimate correct*** (Y/N) 37.6%/62.43%

X – 15 variables (5+ variables where discarded from initial set)

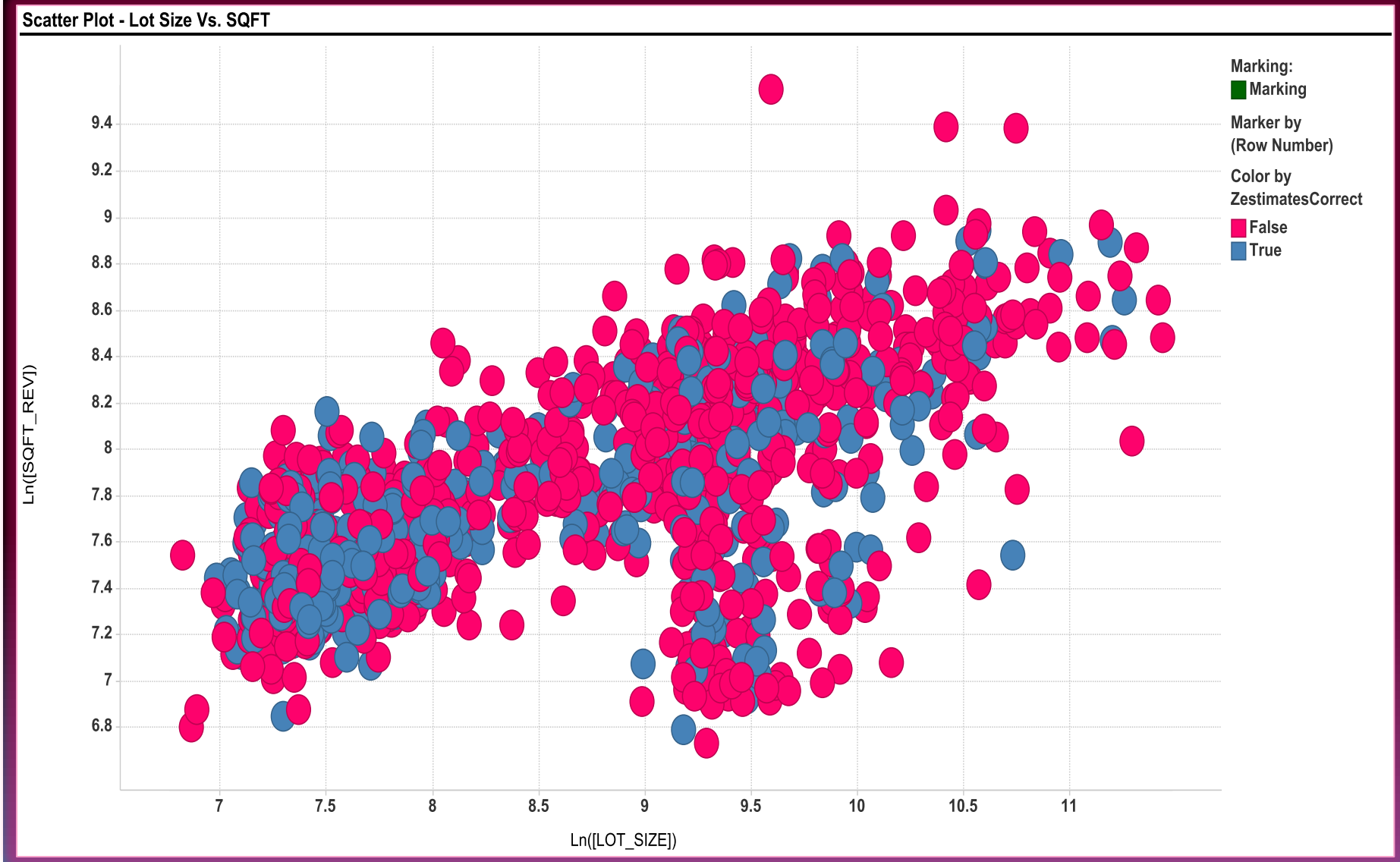
Data Fields – X's

1. Home Type (Single Family, Condo , etc)
2. No of Bed Rooms
3. No of Bath Rooms
4. Total Area – Sqft
5. Lot size – Sqft
6. No of Stories
7. Total Rooms
8. Distance from Metro
9. Primary School Rank
10. Middle School Rank
11. High School Rank
12. Age of house at Sale
13. Sale Season (Fall , Winter , etc)
14. Recession Period (Y/N)
15. Sales Volume

Exploratory Analysis



Exploratory Analysis



Modeling Methods - Classification

Data Exploration – Box Plots , Summary Statistics, Scatter Plots , Bar Charts

Data Preprocessing - Derived Variables , Log of variables.

Classification Tree - Top Variables – Lot Size & Distance from Metro

Principal Component Analysis –

Data Exploration – Scatter Plots and Box Plots

Discriminant Analysis

Logistic Regression - ***2nd Best***

KNN - ***Best***

Data Exploration

Naïve Bayes wasn't used since all our X's are numerical in nature.

Criteria for Model Application & Selection

- Unbiased Perspective → symmetric miss-classification cost
- Lower error rate compared to the Naïve rule applied to the same test data.
- Parsimony of models with equivalent error rates
- Inclusion of Variables (X's) based on Availability and relevance at the time of prediction
- Data Partitioning – Training/Validation/Test – 60:20:20 → KNN and Classification Tree

Top 2 Models

1. KNN

Test Data scoring - Summary Report (for k=11) - KNN			
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	183	39	21.31
TRUE	99	51	51.52
Overall	282	90	**31.91**

2. Logistic Regression

Test Data scoring - Summary Report – Logistic Regression			
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**

	Error Rate	% Improvement
KNN	31.91%	8.6 %
Logistic Regression	32.86%	5.7%

Conclusions

- Marginal but still significant Improvement over the Naïve rule.
- Can be used for further validation by Buyers/Sellers/Real Estate Professional's -> to use Zestimate or re-assess the home value.
- Models are parsimonious and all the inputs are available beforehand
- Road to improvement
  - Get more important variables (X's)
  - Get more data points