Predicting Safety of U.S. Cities

Elyas Akram, Min Chen, Phuong Le,
Bavan Vargese, Fan Zhang

Goal

- Predicting the safety (crime rate) of U.S. Cities

- Crime for a city to be predicted based on
  - population density
  - the number of fast-food chains like McDonalds
  - the number of retail chains like Wal-Mart
  - number of Universities
  - foreclosure and vacancy rate

Data

- Crime data from FBI website for 2007 (grouped by city/state)

http://www.fbi.gov/ucr/cius2007/data/table_08.html

- Data on fast-food chains by city, state for
  - Star Bucks, McDonalds, Burger King and Wendy's

http://www.poi-factory.com/

- Data on number of Universities by City from

http://www.univsource.com/ussc.htm

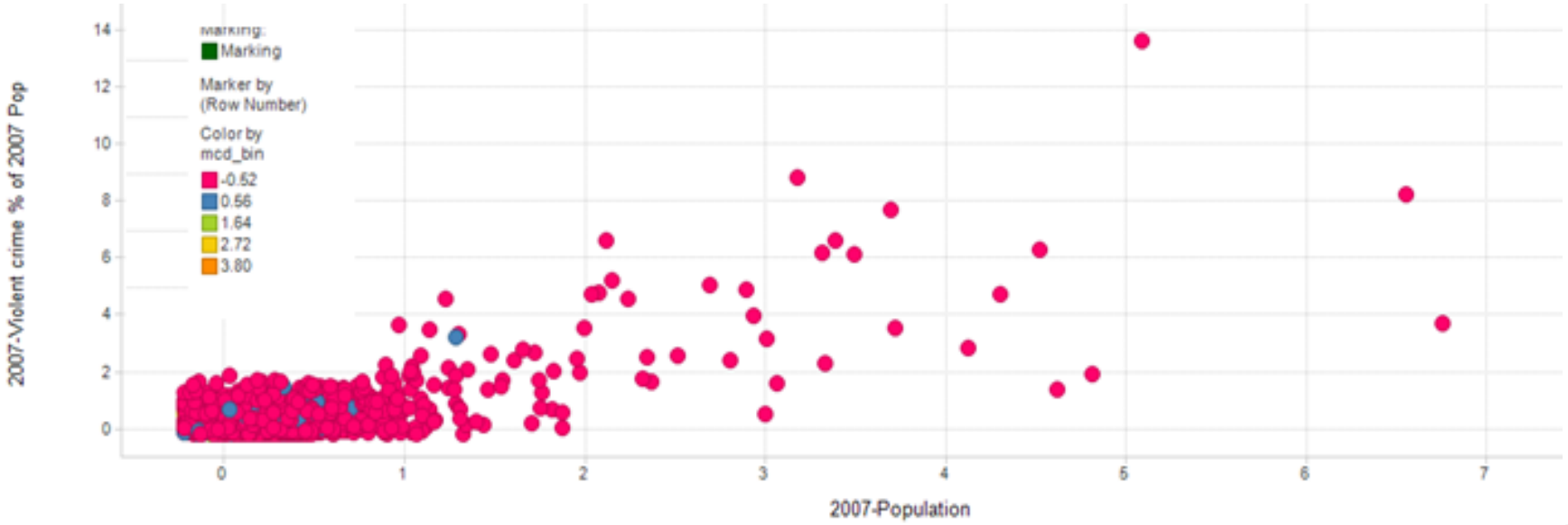
- Foreclosure data from HUD

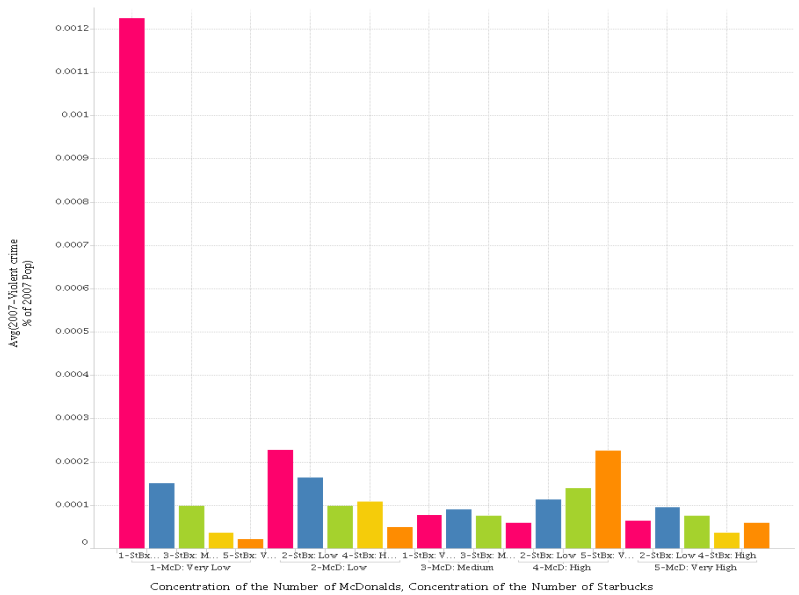
http://www.huduser.org/portal/datasets/pdrdatas.html

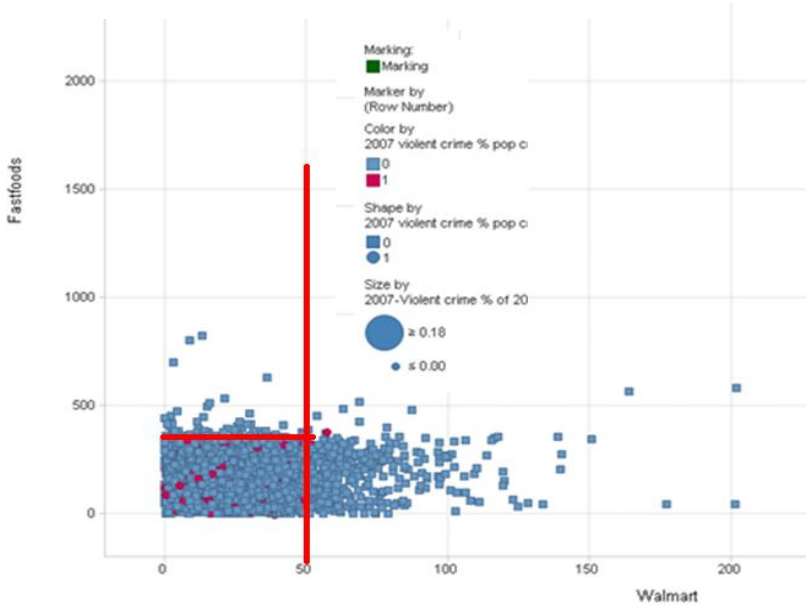
Data Cleanup

- Merged the different data files using City and State as the identifiers
- Initial Data Set Size – 9,379 observations and 37 variables
- Cleaned up data – 3,456 observations and 25 variables
  - Deleted cities with no information on crime

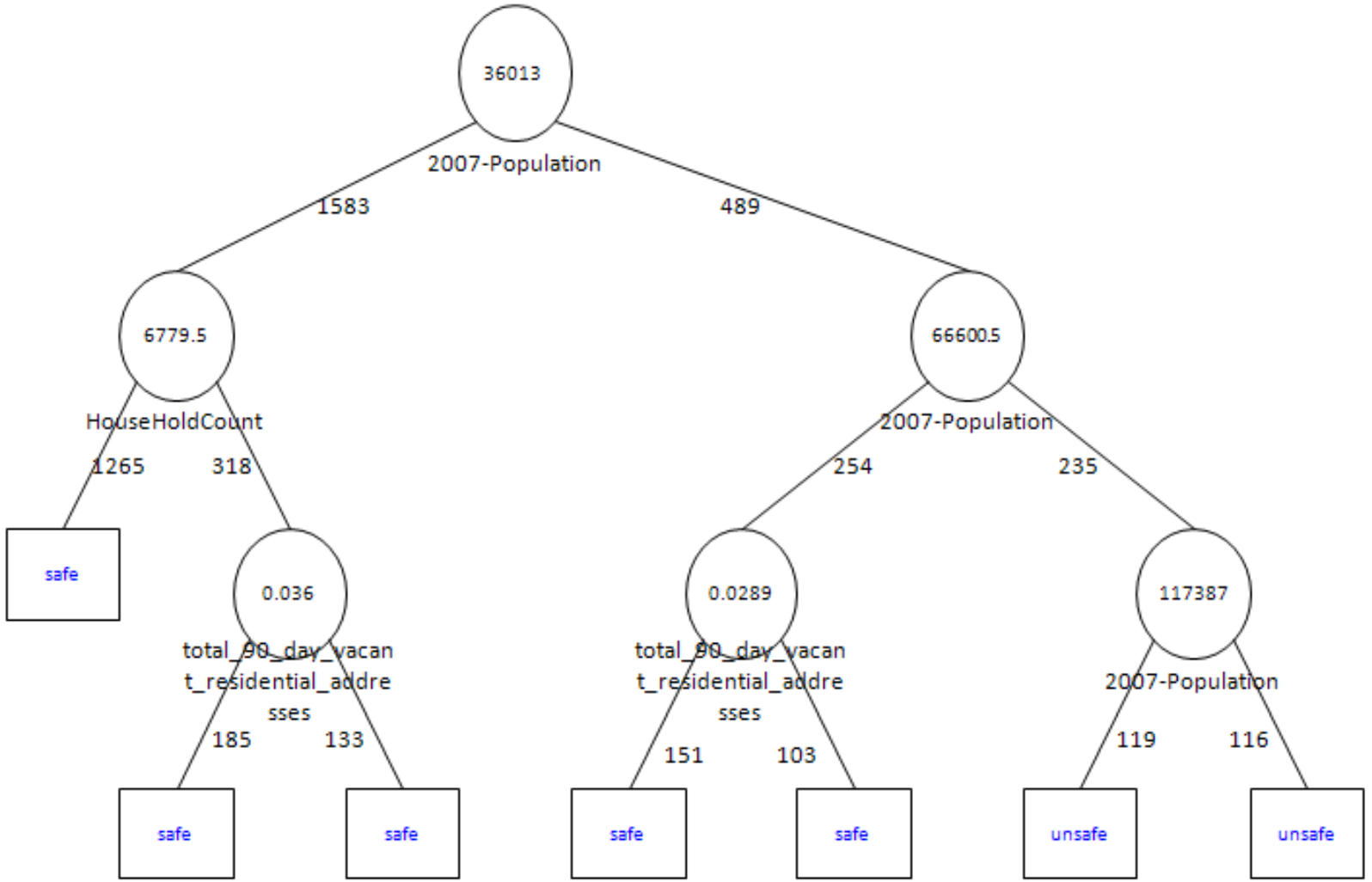
Data Exploration



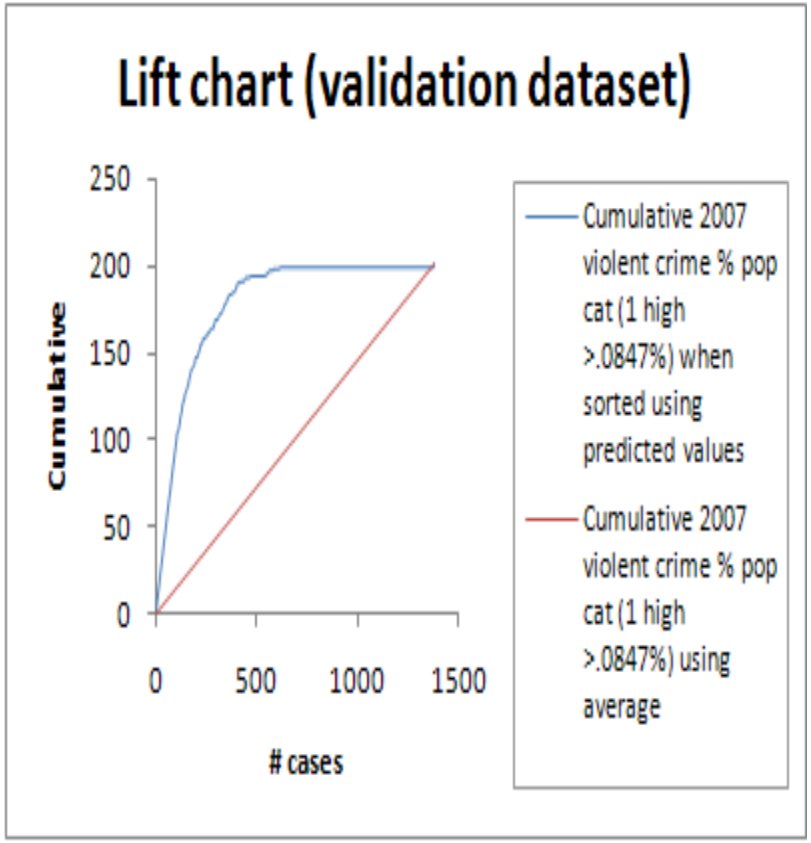




Model 1 – Classification Tree



Model 2 – K Nearest Neighbors



Validation Data scoring - Summary Report (for k=15)

Total sum of squared errors	RMS Error	Average Error
71.57095769	0.227569677	0.001263034

Model 3 – Discriminant Analysis

Variables	Classification Function	
	1	0
Constant	-1.94311512	-0.72962779
estimated_number_foreclosu	0.70414591	-0.1114331
total_90_day_vacant_reside	-0.14286315	-0.00923953
Total FastFood Restaurant	1.5478282	-0.26120687
Total Universities	0.22839473	-0.04210817

Validation Data scoring - Summary Report

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	1	0
1	124	94
0	51	1113

Error Report			
Class	# Cases	# Errors	% Error
1	218	94	43.12
0	1164	51	4.38
Overall	1382	145	10.49

Final Model – Logistic Regression

The Regression Model

Input variables	Coefficien	Std. Error	p-value	Odds
Constant term	-8.080656	0.4921519	0	*
2007-Population	0.0000795	9.19E-06	0	1.0000795
BurgerKingCount	0.1408806	0.0956627	0.1408372	1.1512871
McDonaldsCount	0.0656035	0.0436968	0.133269	1.0678033
StarBucksCount	-0.186635	0.0697379	0.0074455	0.8297468
WalmartCount	0.0449187	0.0799125	0.5740488	1.0459429
WendysCount	0.2604008	0.1120597	0.0201379	1.2974501
Total universities	0.2134254	0.0744697	0.0041578	1.2379111
HouseHoldCount	8.326E-05	2.292E-05	0.0002813	1.0000832
estimated_number_mortga	-0.000187	3.195E-05	0	0.9998126
estimated_number_foreclo	34.213181	4.5963993	0	7.221E+14

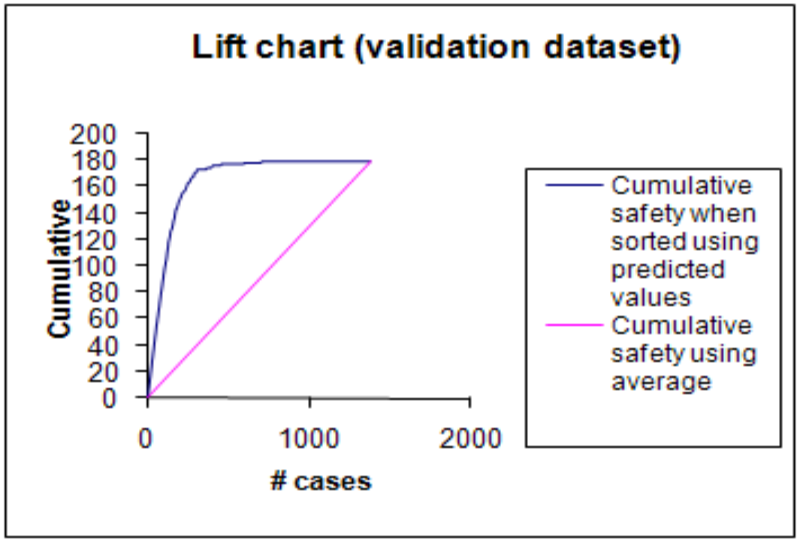
Residual df	2061
Residual Dev.	487.08365
% Success in training	13.465251
# Iterations used	10
Multiple R-squared	0.7025343

Validation Data scoring - Summary Report

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

Classification Confusion Matrix		
	Predicted Class	
Actual Class	unsafe	safe
unsafe	131	48
safe	24	1179

Error Report			
Class	# Cases	# Errors	% Error
unsafe	179	48	26.82
safe	1203	24	2.00
Overall	1382	72	5.21



Conclusion…

- Bad news..
  - Foreclosures are very bad!!
  - Higher learning = higher crime?
  - More fast food restaurants, Wal-Marts indicate an unsafe environment
- But..
  - Starbucks is a star
  - More mortgages are good!

Questions?