City of Syracuse Property Vacancy Project

IST 600
April 26, 2018

Introduction

Main Goal:

To find out what features and variables contribute to or relate to vacant properties in the City of Syracuse

Broad approach:

Decided on the objectives → Collected data → Cleaned data → Merge 3 datasets → Cleaned data → Identified variables to model → Modeling data → Present results

Introduction Cont.

Data description

- 3 Categories of data:

Crime Data (2017) → Vacant Property Data (2017) ← Census Data (2010)

Process of data combination

- Identified block address to match with
- Merged at block level

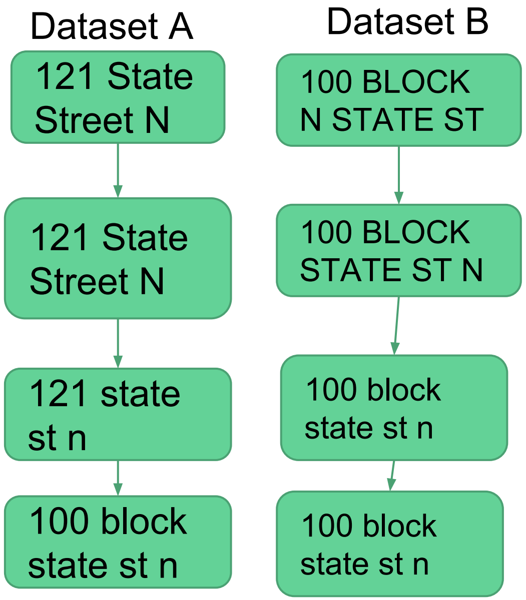
Note: Crime Data referred to dataset A, Vacant Property dataset B, Census dataset A

Merging the 3 Datasets

Step 1: Change the address format(order) to “StrNum StrName St/Av/Rd/Pl Direction”

Step 2: Change the synonym into the same words, eg: “Avenue” “Ave” => “Av”, lower case all address

Step 3: Create block for each property address (1xx -> 100 block), merge them together.



Results:

There are 347 of 2013 blocks with crimes containing no property.

There are 18739 of 42372 properties containing no crime data.

Final dataset format

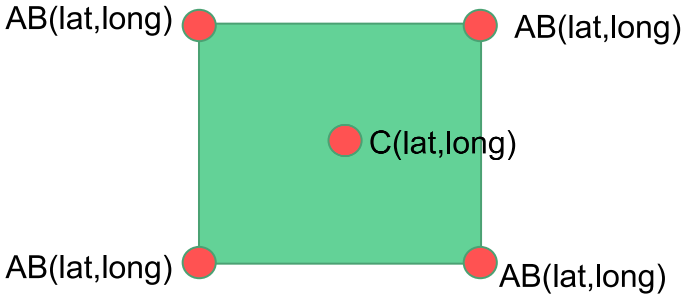
Property address	Block address	Features A	Features B	Features C
121 state st n	100 block state st n	AAAAAA	AAAAAA	AAAAAA
122 state st n	100 block state st n	AAAAAA	AAAAAA	AAAAAA
223 state st n	200 block state st n	AAAAAA	AAAAAA	AAAAAA

Merging the 3 Datasets

1. We then used the A (crime) and B (vacancy) merged dataset at block level and merged with dataset C (Census).

2. Found lat, long for the addresses present in A,B merged data

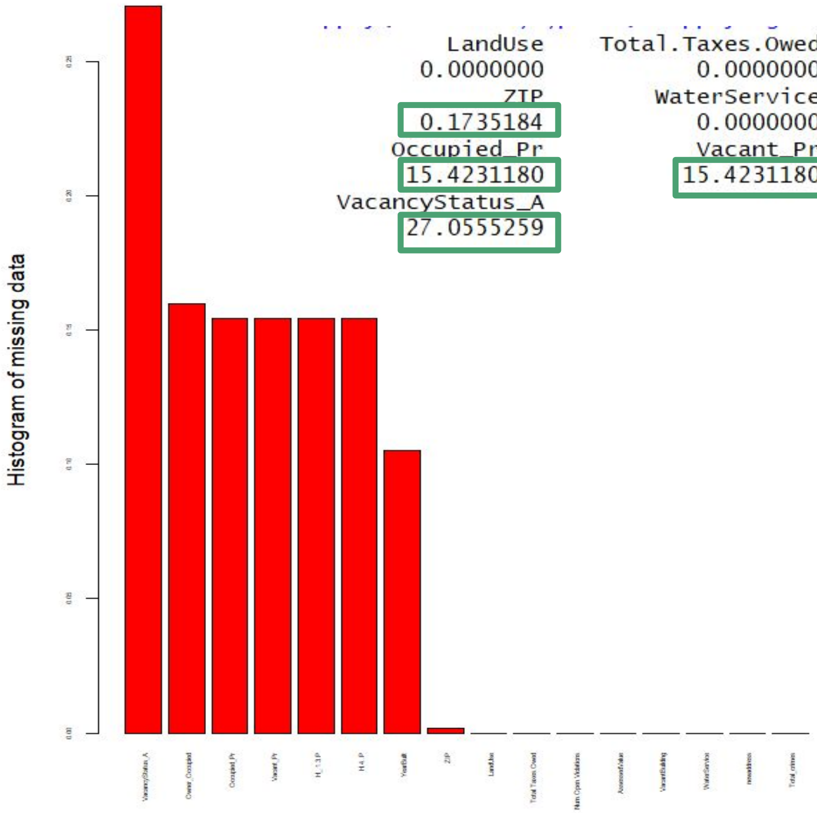
3. Used KNN algorithm to assign lat, long point of dataset C to the nearest lat, long points of dataset AB.



Final dataset format

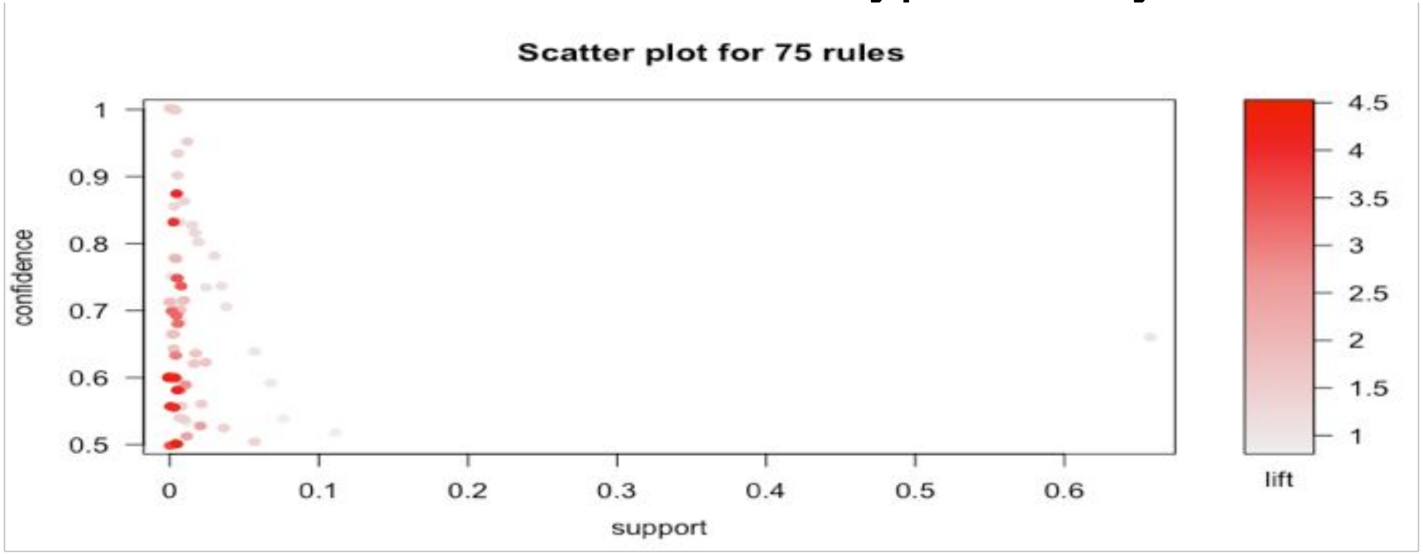
Property address	Block address	Features A	Features B	Features C
121 state st n	100 block state st n	AAAAAA	AAAAAA	AAAAAA
122 state st n	100 block state st n	AAAAAA	AAAAAA	AAAAAA
223 state st n	200 block state st n	AAAAAA	AAAAAA	AAAAAA

Data Cleaning/Preparation



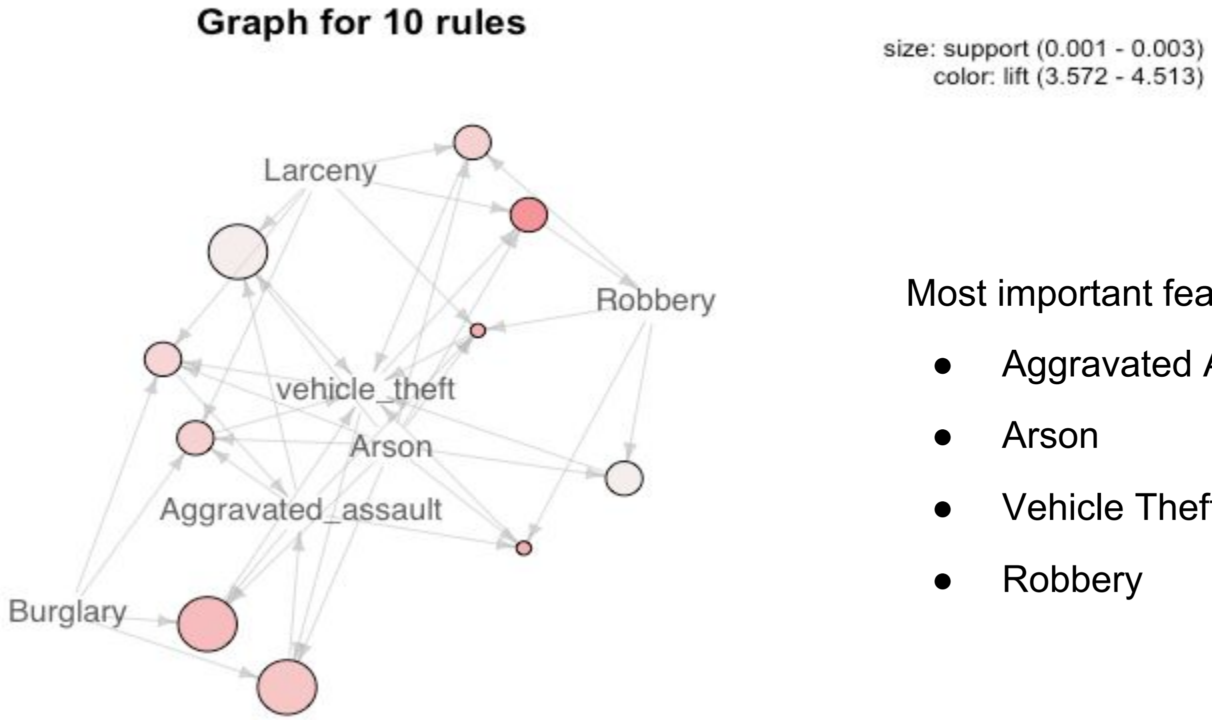

```
          LandUse    Total.Taxes.Owed Num.Open.Violations       AssessedValue     VacantBuilding
        0.0000000           0.0000000           0.0000000           0.0000000          0.0000000
              ZIP        WaterService           YearBuilt          newaddress       Total_crimes
        0.1735184           0.0000000          10.5312333           0.0000000          0.0000000
      Occupied_Pr          Vacant_Pr             H_.1.3.P              H.4..P     Owner_Occupied
       15.4231180          15.4231180          15.4231180          15.4231180         15.9837160
  VacancyStatus_A
       27.0555259
```

Association rules for different crime types in Syracuse



	support	confidence	lift	count
[1] {Arson, Larceny, vehicle_theft} => {Robbery}	0.002472799	0.5000000	4.513393	5
[2] {Aggravated_assault, Arson, Robbery} => {vehicle_theft}	0.001483680	0.6000000	4.286926	3
[3] {Aggravated_assault, Arson, Larceny, Robbery} => {vehicle_theft}	0.001483680	0.6000000	4.286926	3
[4] {Aggravated_assault, Arson, Burglary} => {vehicle_theft}	0.003461919	0.5833333	4.167845	7
[5] {Arson, Burglary, vehicle_theft} => {Aggravated_assault}	0.003461919	0.8750000	4.104988	7

Graphic visualization for A-rules with highest lift



Most important features:

- Aggravated Assault
- Arson
- Vehicle Theft
- Robbery

Selected Columns

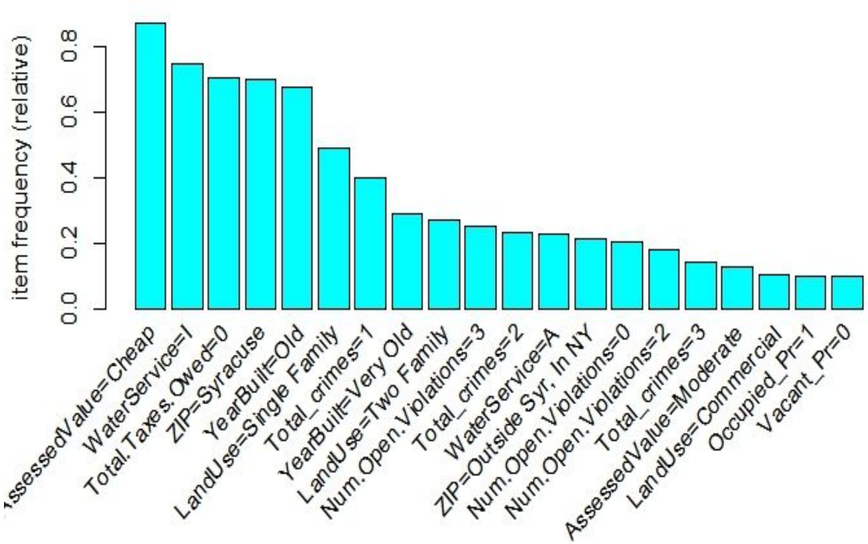
Land use	Occupied probability
Open violations	Households 1-3 occupants
Assessed value	Households 4+ occupants
Vacant building	Aggravated assault
Owner zip code	Arson
Year built	Robbery
Owner occupied	Vehicle theft

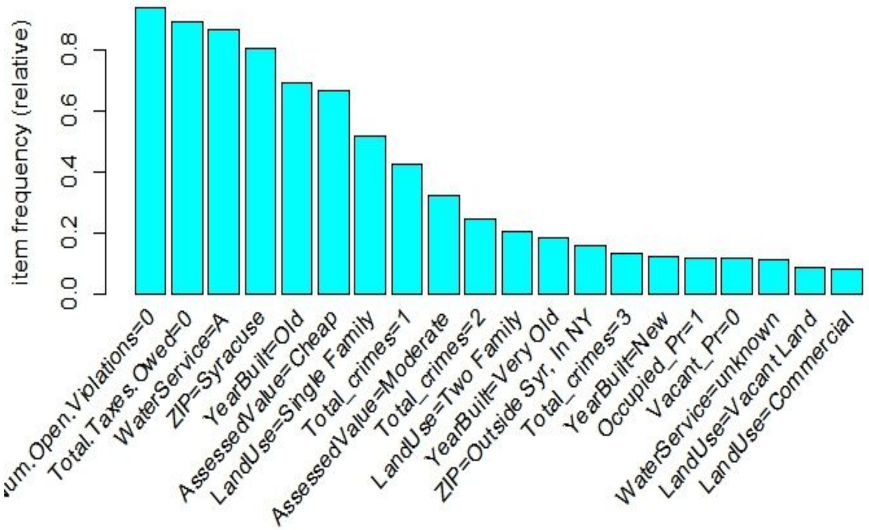
Apriori Rules for Vacant Building Types

```
    lhs                          rhs                  support confidence     lift count
[1] {Num.Open.Violations=3,
     WaterService=I}          => {VacantBuilding=Y} 0.01039780  0.9509202 19.00183   155
[2] {LandUse=Two Family,
     WaterService=I}          => {VacantBuilding=Y} 0.01106863  0.7857143 15.70059   165
[3] {LandUse=Single Family,
     Total.Taxes.Owed=0,
     ZIP=Syracuse,
     WaterService=I}          => {VacantBuilding=Y} 0.01006239  0.7853403 15.69312   150
```

```
    lhs                          rhs                 support confidence     lift count
[1] {Total.Taxes.Owed=0,
     Num.Open.Violations=0,
     ZIP=Syracuse,
     WaterService=A}          => {VacantBuilding=N} 0.5850943  0.9965722 1.049072  8722
[2] {Total.Taxes.Owed=0,
     Num.Open.Violations=0,
     WaterService=A,
     YearBuilt=Old}           => {VacantBuilding=N} 0.5319648  0.9962312 1.048713  7930
[3] {Num.Open.Violations=0,
     WaterService=A,
     YearBuilt=Old}           => {VacantBuilding=N} 0.5730194  0.9961516 1.048629  8542
```

Item Frequency Plot





Models - Naive Bayes

After bucketizing Owner occupied, Number of persons in the households, Total taxes owed, and Total crimes

Naive Bayes was modelled and the model predicted with an accuracy of 85.26%

The confusion matrix for the entire model is shown below:

Prediction Of Vacant Building	No	Yes
No	98	19
Yes	14	93

Models - Naive Bayes

Surprise, Surprise

Feature selection was done using the null to optimum model:

1) Number of open violations alone predicted 84% accurately
2) When Total crimes was added to the model, the model predicted 84.5% accurately
3) Assessed Value and Water Service, when taken alone did not have a good percentage of prediction
4) When performing feature selection felt, people with prior knowledge in the field can put the models to better use by changing features

Models - Logistic Regression

Predictor Variables: Land Use, Number of Open Violations, Assessed Value, ZIP, Water Service and Year Built

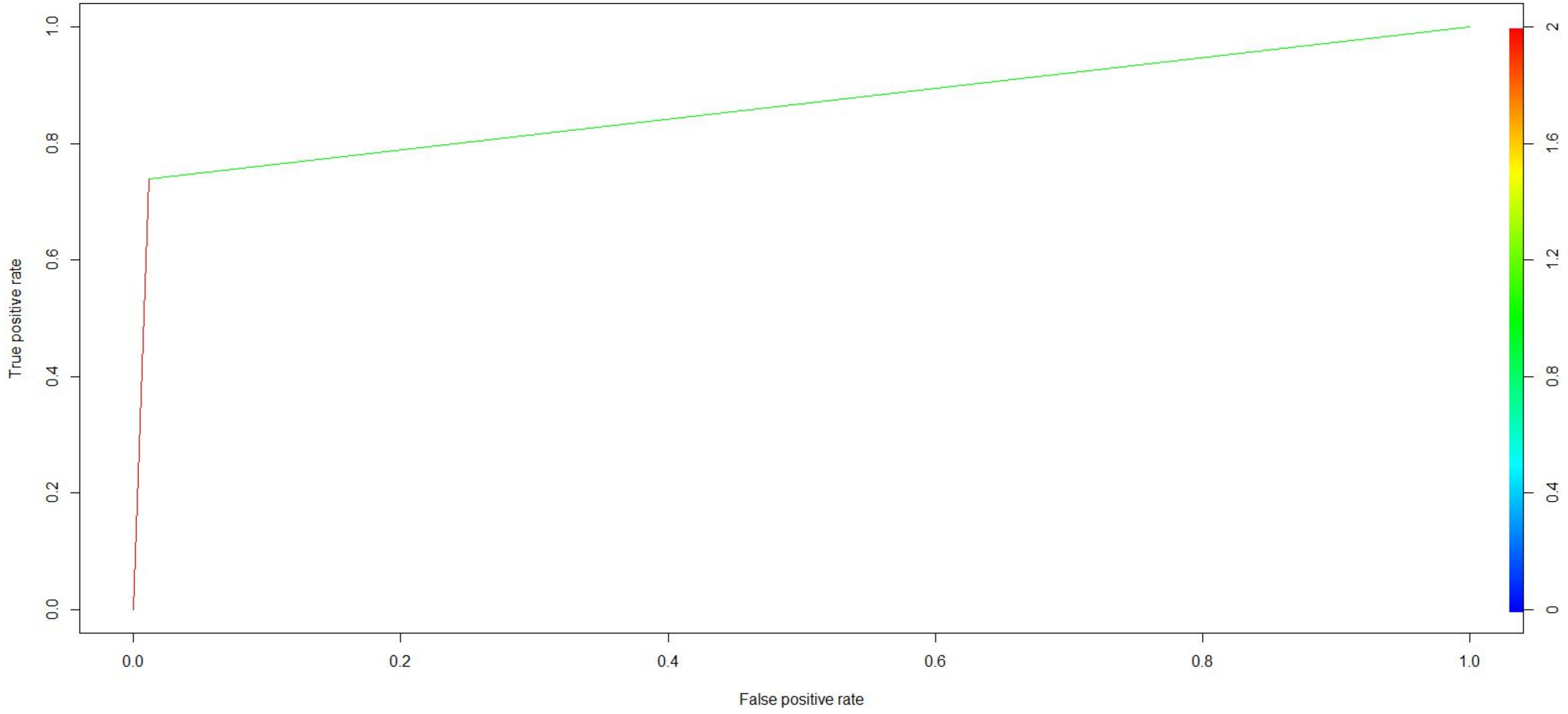
Accuracy: 97.62%

Sensitivity (True Negative rate) - 73.9%

Specificity (True Positive rate) - 98.8%

Prediction Of Vacant Building	No	Yes
No	4674	56
Yes	62	176

Models - Logistic Regression



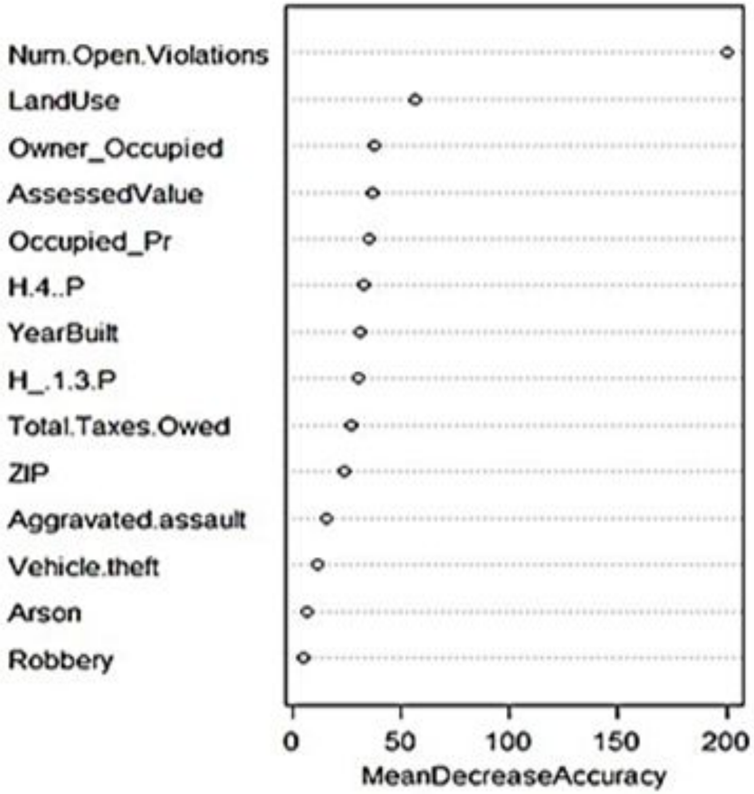
Model - Random Forest

Predictors		
Y	Vacant building	
X	Open violations	Households 1-3 occupants
	Assessed value	Households 4+ occupants
	Land use	Aggravated assault
	Owner zip code	Arson
	Year built	Robbery
	Owner occupied	Vehicle theft
	Total Tax Owed	
	Occupied probability	

- Sample – 14985
- Accuracy – 98.65%
- Confusion Matrix

		Actual	
Prediction		N	Y
	N	14187	**165**
	Y	**35**	583

Key Predictors



Model - Support Vector Machines

Predictors		
Y	Vacant building	
X	Open violations	Households 1-3 occupants
	Assessed value	Households 4+ occupants
	Land use	Aggravated assault
	Owner zip code	Arson
	Year built	Robbery
	Owner occupied	Vehicle theft
	Total Tax Owed	
	Occupied probability	

- Sample – 14985
- Accuracy – 96.42%
- Confusion Matrix

		Actual	
Prediction		N	Y
	N	14111	**411**
	Y	**125**	337

Model – K Support Vector Machines

Predictors		
Y	Vacant building	
X	Open violations	Households 1-3 occupants
	Assessed value	Households 4+ occupants
	Land use	Aggravated assault
	Owner zip code	Arson
	Year built	Robbery
	Owner occupied	Vehicle theft
	Total Tax Owed	
	Occupied probability	

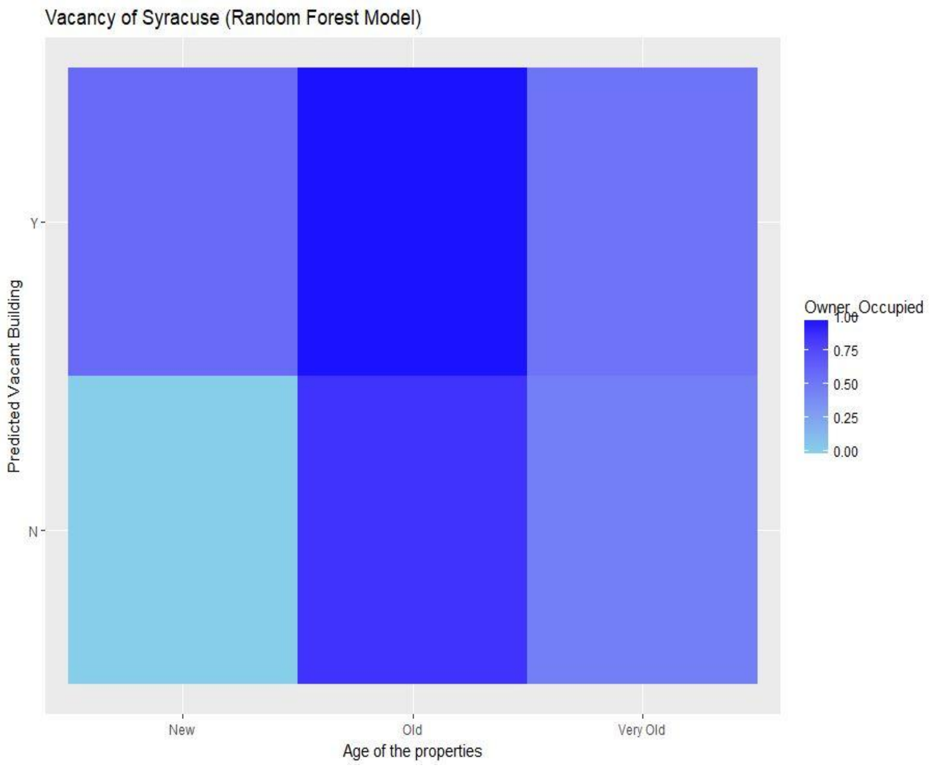
- Sample – 14985
- Accuracy – 97.48%
- Confusion Matrix

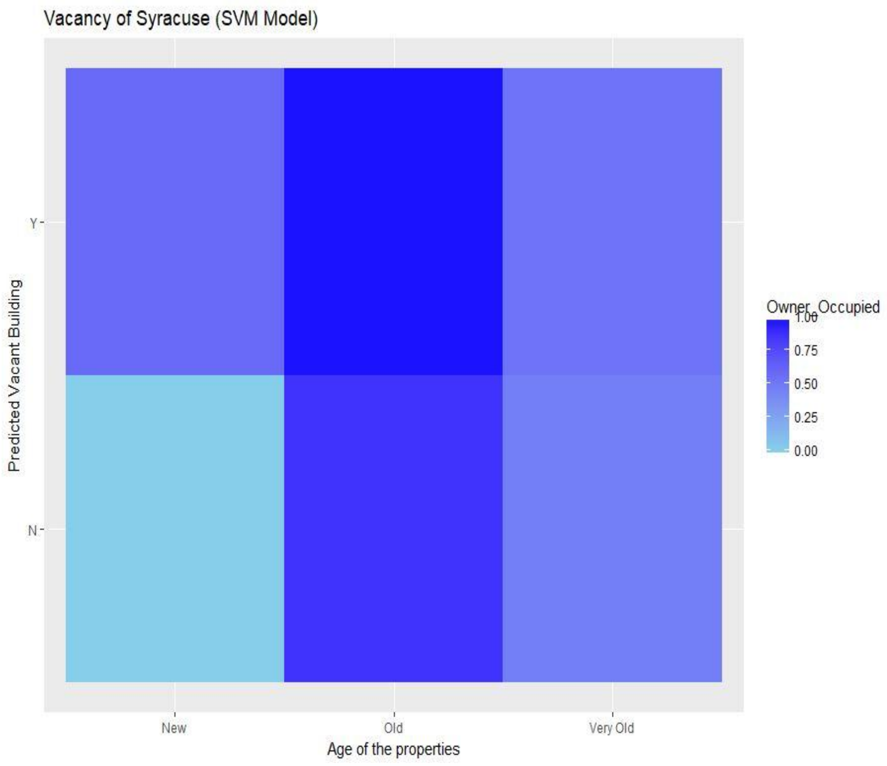
	Actual		
Prediction		N	Y
	N	14105	**246**
	Y	**131**	502

Results

Models	**Vacant Building (Yes)**	**Vacant Building (No)**
Random Forest	618	14367
SVM	**462**	14523
ksvm	633	14352

Predicted vacancy (Vacant Building) based on condition and ownership





Interpretation of Results

RF/SVM - More the number of open violations higher the probability of land being vacant. Landuse is another important predictor.

Apriori - If number of open violations are more than 2 and water services are inactive, higher is the probability of land being vacant.

Interpretation of Results

Logistic - When the Assessed Value of the property is moderate i.e. between the price range of $75000 and $2000000, there are higher chances of the property being vacant.

Odds for very old buildings (before 1900’s) to be vacant is 139% higher than odds of a new building (1976 - 2017) being vacant.

Odds for old buildings (1975 - 1900’s) to be vacant is 120% higher than odds of a new building (1976 - 2017) being vacant.

For a unit increase in open violations, there is an 18% increase in the odds of a building being vacant.

Recommendations

1. We have only used 2010 Census data but it would interesting to include income data which is available from the American Community Survey (ACS).
2. Run different models using the same set of data variables to ensure results show some sort of prediction consistency.

Note: We have removed I-81 data (Syracuse Highway) from the 2010 Census downloaded since the dataset provides several addresses from the highway that were not useful for our analysis.

Appendix

Data repository

Code Reusability