5-2017

A Machine Learning Approach on Providing Recommendations for the Vacant Lot Problem

Md Towhidul Absar Chowdhury
mac9908@rit.edu

A Machine Learning Approach on Providing Recommendations for the Vacant Lot Problem

by

Md Towhidul Absar Chowdhury

A Thesis Submitted in Partial Fulfillment of the Requirements for the
Degree of Master of Science
in Software Engineering

Supervised by

Dr. Naveen Sharma
Department of Software Engineering
B. Thomas Golisano College of Computing and Information Sciences
Rochester Institute of Technology
Rochester, New York
May 2017

The thesis “A Machine Learning Approach on Providing Recommendations for the Vacant Lot Problem” by Md Towhidul Absar Chowdhury has been examined and approved by the following Examination Committee:

Dr. Naveen Sharma, Professor Date
Thesis Committee Chair and Advisor

Dr. Pradeep Murukannaiah, Assistant Professor Date
Committee Member

Dr. Scott Hawker, Associate Professor Date
SE Graduate Program Director

Dedication

To my parents, who instilled in me a love for the academia. To my sister, who always inspired me to follow in her footsteps and excel.

Acknowledgments

I would like to thank my committee members Naveen Sharma and Pradeep Murukannaiah for helping and supporting me through this research. Special thanks to my advisor Naveen Sharma for pushing me forward in the right direction, and guiding me through my Masters journey.

I would also like to thank my roommate Naseef Mansoor, our conversations and discussions about research and problem solving really provided a catalyst for my journey.

Last but not the least, I would like to thank my parents, my sister and all my friends. Without your support, it would have been impossible to succeed.

Abstract

A Machine Learning Approach on Providing Recommendations for the Vacant Lot Problem

Md Towhidul Absar Chowdhury

Modeling municipal or urban decisions is challenging due to the abundance of variables that guide end results. One such challenging issue is the existence of vacant lots in a city, which causes poorer standard of living for the community. As a result, reclaiming these properties and putting them into productive use is a primary concern. However, each time community leaders had to “reinvent the wheel” and make decisions from scratch. To this end, we propose the creation of a vacant lot model and utilizing it to provide recommendations for vacant lot conversions, providing a starting point for such decision making. We define a vacant lot model in terms of determinants to a vacant lot’s impact, and evaluate the proposed method on real-world vacant lot datasets from the cities of Philadelphia, Pennsylvania and Baltimore, Maryland. Our results indicate that our prediction model performs accurately on cities with a centralized approach to vacant lot conversion.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Background

We live in a data driven society, where every decision and every plan has to be supported with historical data or statistics. Established data sets, such as Data.gov [2], provide ample data for macro-level analysis and decisions. However, at the community or neighborhood level, such as monitoring urban issues like traffic congestion, poverty or health care, these datasets by themselves are not sufficient.

In the context of urban planning, a common issue for most cities are the existence of vacant lots. A vacant lot is an abandoned property that has no buildings on it [5]. In the past these properties did have buildings or houses on them but they were demolished due to safety concerns as they became abandoned and fell into disrepair.

These vacant lots are an issue of concern because they have a tendency to attract illegal activities such as littering or dumping of solid waste, or even spaces where criminal activities may take root. Since vacant lots result in a poorer standard of living for the urban community [14], reclaiming these properties and putting them into productive use has become a primary concern for the community.

In order to do that, the primary goal is to fill the empty space with attractive and productive activities or projects that would not only bolster the people's attention but also increase the health of the community overall. For example, the vacant lots can be converted to temporary community gardens, urban farms etc. But the conversion of all vacant lots is not feasible, as urban planners and community leaders have to focus on optimizing their decision making. They need to ascertain which lots will provide the most benefit once they are converted. Furthermore, vacant lots are a liability [5] that cause depressed property value in the surrounding neighborhood [8]. As a result, urban planners may also focus on converting lots to prevent the negative effect on neighborhood property value.

In their effort to tackle vacant lot conversions, urban planners and community leaders have to routinely analyze available data, and prioritize zones. They need to identify how similar vacant lots in the past have been converted, and the resulting effects of those conversions. Each time a vacant lot needs to be converted, they have to "reinvent the wheel" and start from the beginning.

While much research has been done in analyzing the impact of converting vacant lots [8, 12, 10], there is little research on predicting which vacant lots to convert based on the determinants of previously successful vacant lot conversions. Furthermore, while significant amount of data is available from cities that have tackled or are facing the problem, there is a lack of normalized vacant lot data that would enable stakeholders to analyze and estimate the optimality of the vacant lots.

The goal of this research is twofold: 1. develop general datasets consisting of existing vacant lots and converted lots consisting of possible determinants of vacant lot conversion, and 2. from this dataset, develop and analyze sets of prediction models to predict which vacant lots should be converted, and establish a foundation for further research in solving similar problems.

The motivation behind this thesis is discussed in Section 1.2. Section 1.3 outlines the specific research objectives for this project. Chapter 2 reviews related works in current literature, Chapter 3 describes the approach taken in analyzing and solving this problem. Chapter 4 discusses the results from our experiments and Chapter 5 provides a summary of the work done and the results, and outlines future work.

1.2 Motivation

At present, the primary method for tackling the vacant lot problem are programs that attempt to engage members of the community to actively work towards converting them. Community outreach programs such as “Grounded in Philly” in Philadelphia, Pennsylvania [3], or the “Adopt-A-Lot” [1] program in Baltimore, Maryland provide information for the general public to assist in communities converting vacant lots.

But such programs only provide the information and data, and as a result there is a barrier to entry for anyone wishing to convert a vacant lot. A system that would automatically parse the data and provide a recommendation to users on which lots would provide the most benefit would go a long way towards removing that barrier.

1.3 Research Objectives

There are two objectives that this thesis focuses on. The first objective is focused towards building a general dataset that can be used to identify the determinants of a vacant lot conversion in target cities.

The second is to utilize the dataset to determine if a prediction model can be made for vacant lot conversions, evaluate the prediction model and provide foundation for a system that would ease the burden of vacant lot selection on community leaders. The end result of this research objective is to present preliminary predictions on which vacant lots should be converted, analyze why these vacant lots were chosen and to provide a starting point for urban planners to focus limited resources to prioritize certain vacant lots. It is important to note that the purpose of this research is *not* to develop a new novel algorithm for predicting vacant lot conversions, but to use existing models in literature to build a vacant lot recommender.

Chapter 2

Related Works

The concept of vacant lots and possible solutions are modeled extensively well in existing literature. Accordino et. al. [5] provides a detailed look at how the existence of vacant lots cause concern for the community, and provides an overview of how they are solved in various cities. Accordino [5] further concludes that the solution to such problems not only happen from an urban planning side, but also from the neighborhood community as well. However, the study is focused towards cities as a whole and does not take into account whether each problem tackled resulted in success for those vacant lots.

A better look at the effects of vacant lot conversion is covered by Branas et. al. [7], where estimates showed that vacant lot greening was associated with consistent reductions in gun assaults across all four sections of the city and consistent reductions in vandalism in one section of the city. Regression-adjusted estimates also showed that vacant lot greening was associated with residents reporting less stress and more exercise in select sections of the city. Once greened, vacant lots may reduce certain crimes and promote some aspects of health.

Furthermore Kremer [14] also suggests that by assessing vacant lot uses, ecological characteristics and the social characteristics of neighborhoods in which vacant lots are located, urban planners may be able to more effectively address vacant lots while promoting urban sustainability and resilience. Automating such analysis using machine learning algorithms may decrease effort for urban planners.

Similar systems for making suggestions have been developed, and their use in different types of data are highlighted in Capdevila et. al. [9] for geolocation based data and in Ramesh et. al. [4] for social media data. These papers, however, provide a recommender system approach rather than a machine learning approach.

Tayebi et. al. [16] presented a novel approach to crime suspect recommendation utilizing a random walk method based on partial knowledge of offenders involved in a crime incident and a known co-offending network.

Ruining et. al. [11] built a large-scale recommender systems to model the dynamics of a vibrant digital art community, Behance, consisting of tens of millions of interactions (clicks and 'appreciates') of users toward digital art.

Our prediction model needs to estimate the present utility of the vacant lot along with the future utility after a conversion. The concept of utility for urban infrastructures, including vacant lots, was introduced in [15], along with a detailed case study analysis. The theory and data analysis tools provided by the paper can be utilized in solving similar problems, but most of the heavy burden of performing the calculations fall on the data scientist or urban planner. The introduction of the proposed system will remove this burden and increase planning efficiency.

Chapter 3

Approach

3.1 Determinants of A Vacant Lot Conversion

This section proposes a unified formal model of describing a vacant lot in terms of attributes related to its surrounding neighborhood. This model is built on the assumption that each vacant lot has a set of features that define the impact converting the particular vacant lot will have. Specifically, the formal model aims at bridging the conceptual gap between data level, mining level and interpretation level, and facilitates separating the description of data from the details of data mining and analysis. By gradually transforming and reducing the unified model to more specific views, we obtain the final vacant lot model as one such view.

For the purposes of our prediction model and the classification of whether a vacant lot should be converted. A set of vacant lot, $V = \{v_1, v_2, v_3, ..., V_n\}$, consists of all vacant lots in the dataset with each vacant lot, V_i represented as a dependency of its feature set F where:

$$F = \{f_1, f_2, f_3, f_4, f_5, f_6\} \tag{3.1}$$

f_1 = Utility from public services and infrastructure

f_2 = Access to vacant lot

f_3 = neighborhood property value indicator

f_4 = vacant lot density

f_5 = crime density

f_6 = zone

Detailed discussion on the selection of each of the attributes, how they were collected and how they are represented in our final datasets are discussed in detail in the following sections.

3.1.1 Utility from Public Services and Infrastructure

A primary indicator of the wellness of any land unit is the utility received from the closest public infrastructure. As described in Meidar-Alfi [15], distance from these infrastructure is inversely proportional to the utility the target location receives. Public infrastructure common to all cities are public libraries, schools and parks.

In order to get an indication of utilities provided by these facilities, for each vacant lot we calculated the distance in meters to the closest library,park and school. An increasing distance from the closest public infrastructure would result in a decreased utility received from those infrastructures.

To normalize distances across different cities, and to account for geographic and spatial weights, the distances were split into ten equal portions and assigned a score from 1-10, with 10 indicating the lowest distances and

1 indicating the highest distances. This normalization method is used to convert pure distance measures into categorical indicators known as utility scores. A utility score indicates the amount of utility or benefit received from an urban infrastructure as defined by Meidar-Alfi in [15]. As a result, our model will focus more on relative distances across cities rather than overfit on the exact distances.

After the necessary calculations, we get three attributes *distance to school*, *distance to park*, *distance to library*. They were combined into a public utility score as given in Equation 3.2.

$$U = [(S_{w_i} + P_{w_i} + L_{w_i})/3] \tag{3.2}$$

U = Total utility score

S = Utility score from school

P = Utility score from park

L = Utility score from library

w_i = Weight of each utility

3.1.2 Access to Vacant Lots

Once a vacant lot is converted to a green space, the benefit it provides will be dependent on the ease of access it has. A study by Wachter et. al. [17] analyzes the effect of public transit on vacant land management, and suggests that it may be a determinant.

Similar to our measure of public infrastructure utility, we also measure distance from each vacant lot to the nearest public transit stop, with the data having been normalized to utility scores after the distances were calculated.

3.1.3 Neighborhood Property Value

There is a significant amount of work in current literature that focuses on the impact of vacant lots on neighborhood property values. Most of the work done focuses on a hedonic or spatial difference-in-difference analysis of the impacts [12, 8, 7]. We utilize concepts from both these approaches in estimating how a vacant lot or community garden affects the neighborhood property values at present time.

Utilizing property assessment data for target cities, we calculate the mean property value in a quarter mile radius for each vacant lot for two points in time. For the purpose of our research, we chose property assessment data for the years 2015 and 2014 due to the availability of recent data. The difference between the two points provides a simplified estimate of the trend in property values, and may indicate how the immediate surrounding area is affected by the existence of vacant lots.

A better estimate would have been to compare property values before a vacant lot was converted to a community garden, but due to the unavailability of such data a much more simplified estimate was used. Another indication of the status of the area a vacant lot is situated in is the median property values for that area. In most cases, vacant lots situated in a higher value market have a higher probability of success [7] than those in more distressed market. Furthermore, the existence of vacant lots will negatively affect the overall market value of a neighborhood in the long term.

3.1.4 Vacant Lot Density

In terms of spatial characteristics that define a vacant lot, a primary indicator is to get an understanding of how many vacant lots are in the vicinity i.e. whether there is a cluster of vacant lots in the location. As a result, vacant lot density is calculated. Vacant lot density for each individual vacant lot is defined as the number of vacant lots in a quarter mile radius.

Areas of lower vacant lot density may increase the impact of a vacant lot being converted, while with higher vacant lot density one vacant lot conversion may not result in a significant impact [15, 14, 7].

3.1.5 Crime

An effect on crime through the conversion of vacant lot has been studied extensively in [10]. Due to the extent of the impact on crime, crime density can work as an indicator of the optimality of the vacant lot conversion. However, crime density will vary based on neighborhood population and recommendation models for crime is out of scope for this paper as they have been covered in detail by Tayebi [16].

As a result, we will focus primarily on the number of crime incidents in one standard year around a quarter mile radius for a particular vacant lot.

This gives us a categorical indication of crime patterns for each distinct vacant lot without making our model dependent on representing crime patterns directly.

3.1.6 Zoning Policies

Zoning is the process of dividing land in a municipality into zones (e.g. residential, industrial) in which certain land uses are permitted or prohibited. Thus, zoning is a technique of land-use planning as a tool of urban planning used by local governments in most developed countries.

Every city divides its land into zones with a specific purpose. Each zone defines what can and cannot be built upon the vacant lot, or whether it can be converted as well [6]. Since zoning policies dictate the development of vacant lots so strongly, we used it as an attribute for our vacant lot model.

Zoning policies are categorical variables that are either residential, industrial, business or special purpose.

The resulting model for a vacant lot is summarized in Table 3.1.

3.2 Datasets

This section discusses the datasets used to build the proposed vacant lot model for our experiments. The two primary datasets this research utilizes are from the cities of Baltimore, Maryland and Philadelphia, Pennsylvania. The primary reason for the choice of these two cities was the availability of sufficient quantifiable data on the status of current vacant lots, and also

Variable	**Description**	**Type**
libDist	Distance from the closest library	numeric
parkDist	Distance from the closest park	numeric
schoolDist	Distance from the closest school	numeric
transitDist	Distance from the closest transit stop	numeric
priceDiff	Difference in mean property value	numeric
vacantDensity	Density of vacant lots	numeric
crimeDensity	Density of crime incidents	numeric
zone	Zoning policy for vacant lot	categorical
Target Variable	Status of the vacant lot	

Table 3.1: Summary of the vacant lot model

their determinants. A sample of our dataset from the city of Philadelphia is shown in Figure 3.1.

For public utility measurements and access, we collected map data on library, schools, parks and public transit stops in the city as shown in Figure 3.2. Publicly available crime incident reports were used for calculation of crime density as shown in Figure 3.3, and neighborhood property value data were gathered from yearly assessment records.

The dataset proportion is given in Table 3.2, with the size of each dataset, proportion of vacant lots and their conversions. After all the data was collected, a final dataset was built to represent the vacant lot model. A sample of the dataset is given in Table 3.2.

Furthermore, we also utilize another dataset from the City of Rochester to visually analyze the results from our prediction model and provide recommendations on approaching the vacant lot problem for the particular city.

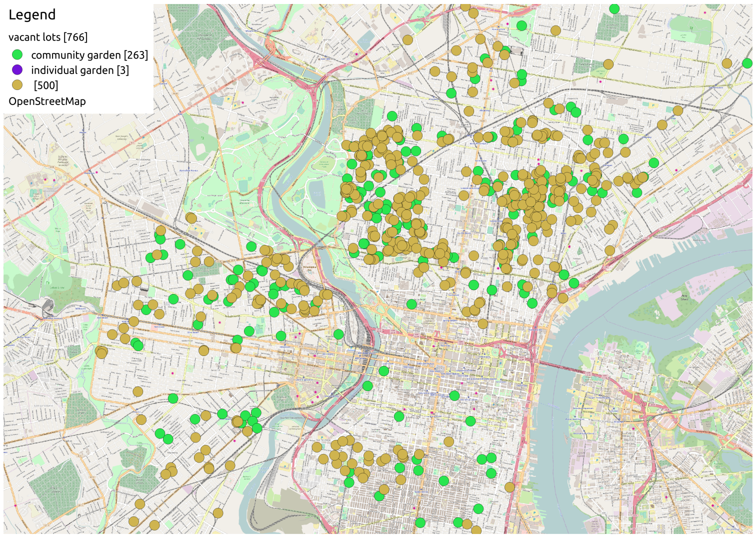


Figure 3.1: Vacant lots and community gardens in Philadelphia

City	Size	Vacant	Conversions
Baltimore	1907	1000	ADOPTED: 517 URBAN FARM: 127 QCMOS: 243
Philadelphia	1101	500	COMMUNITY GARDENS: 601

Table 3.2: Summary of the dataset

3.3 Prediction Models and Classifiers for Vacant Lot Model

We utilized five different classifiers to build our prediction models. The primary reason for this was to analyze the results and the accuracy given by each, as each classifier has their own nuances and tuning parameters to use. We selected each of the classifiers to represent a broader category each of these algorithms fall under. For example, Multilayer Perceptron falls under the broad category of neural network, while k-Nearest Neighbors fall under Instance Based Learning. Each of the classifiers are introduced and a brief description of them are given in the following sections.

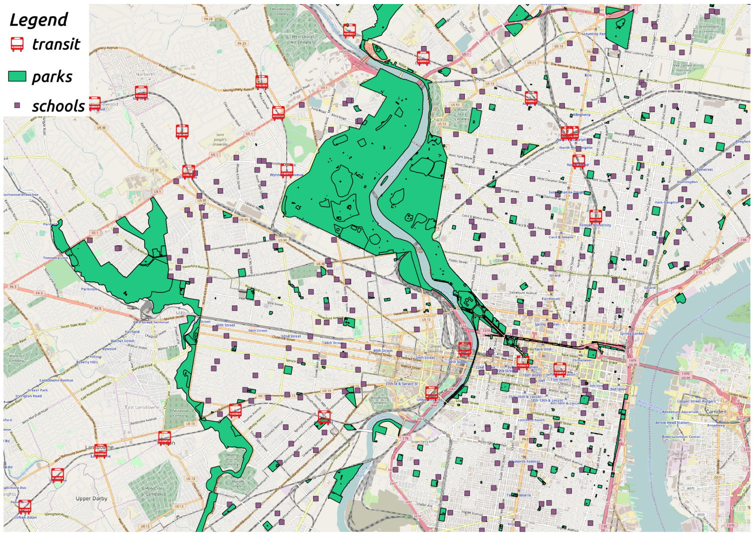


Figure 3.2: Location of public infrastructures in Philadelphia

publicUtil	vacantDensity	crimeDensity	transitDist	priceDiff	category	class
8	36	42	10	1077	R	ADOPTED
4	237	64	10	3026	B	QCMOS
9	147	74	10	-1872	S	URBAN FARM
7	53	91	20	-3315	M	AVAILABLE

Table 3.3: Sample dataset of the vacant lot model for Baltimore

3.3.1 Random Forest

Random forests are an ensemble learning algorithm used for classification problems. Each random forest consists of multiple decision trees that are constructed at training time and classifying based on the each of the attributes in the data. Random forest was developed to tackle the problem of decision trees overfitting to the training data [13].

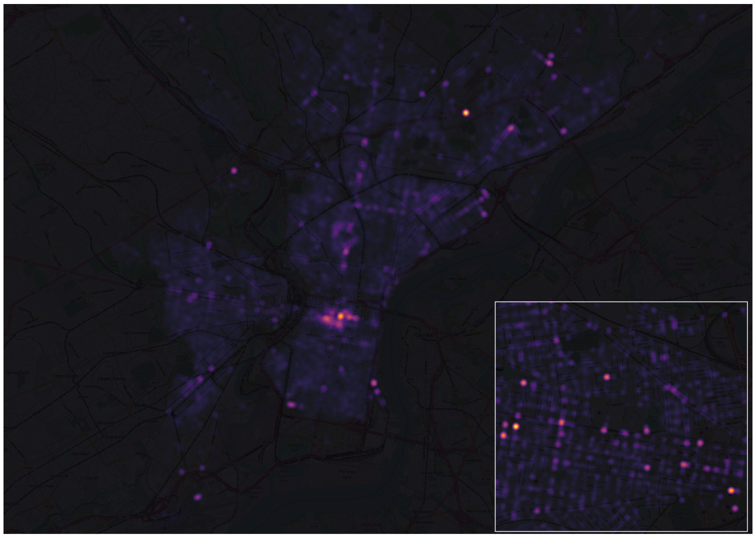


Figure 3.3: Crime density from the Philadelphia dataset

3.3.2 Multilayer Perceptron

A multilayer perceptron (MLP) is a feed-forward artificial neural network. It consists of input nodes, multiple layers of hidden nodes. Each layer is connected to the next layer, and the network itself is represented as a directed graph.

Each node is responsible for processing the input data with the help of an activation function. With each iteration, the network is trained with a backpropagation algorithm that enables weights to be updated with each training instance coming in, to decrease the error of predictions made.

3.3.3 Naive Bayes

Naive Bayes is a classification algorithm based on Bayes Theorem that assumes that each of the independent variables are independent of one another.

It's the simplest form of a Bayesian classifier, and it's strength with categorical variables suit the design of the vacant lot problem as well.

3.3.4 k-Nearest Neighbors

k-NN is categorized as a lazy learner, and falls under the class of instance based learners. It utilizes similarity between objects and an unknown object is classified by a majority vote of its most similar objects. Furthermore, k-NN does not build a model and only approximates a prediction upon receiving an unknown instance.

3.3.5 Support Vector Machines

SVM is a classification technique that constructs linear separating hyperplanes in high-dimensional vector space to separate data points based on their features. The purpose of an SVM is to maximize the separation of the data points from these hyperplanes in order to increase confidence of the classification.

3.4 Experiments

This section describes the methodology and process we followed in experimenting with the generated dataset described in Section 3.2. The objective for these experiments were to build multiple prediction models on different combinations of our pre-determined independent variables, evaluate the accuracy of our classifiers and select the best one that can be used in the future.

We decided to at first analyze how each of the features can be used to predict vacant lot conversion. To that end we built simple classifiers on a subset of features and evaluated their results based on our test set.

3.4.1 Single City Prediction Model

At first, we decided to build a prediction model focused on a single city and study if vacant lot conversions can be predicted within a single city. For our experiments we created five prediction models each for the cities of Baltimore and Philadelphia, with hyperparameters tuned to optimize results on the training set for each city.

For our training set in building those classifiers, our training set was a random sample of 60% of the primary dataset. The remaining 40% of the datasets were used to evaluate the accuracy of our model, and ensure the models provide valid and sane predictions. Furthermore, we built models using both raw data points and normalized utility scores defined in 3.1 to evaluate results from both sets of data and presented the most optimal results.

The classifiers were trained on the training set using a 5-fold cross validation to prevent the models from overfitting to the training data. The results are discussed in detail in Chapter 4.

3.4.2 Cross-City Prediction Model

After analyzing the prediction models for each city, we picked the best classification algorithm and applied them for predictions across cities. We trained our classifiers on Baltimore and tested it on Philadelphia, and vice versa. However, the two test cities in our experiments have different classes that vacant lots were converted to. For example, in Baltimore vacant lots were converted to Qualified Community Managed Open Space (QCMOS), urban farms or simply adopted to a community garden. But in Philadelphia our dataset only consists of community garden conversions. As a result, both dependent variables were changed to a binary class indicating whether a vacant lot has been converted or not.

After the dataset has been updated, similar experiments as described in the previous section was carried out. However, in this case we utilized Baltimore as our training set and validated the results with the datasets from Philadelphia, and vice versa. We did not experiment with raw values for our cross-city prediction model as the raw values will only be useful for particular cities and not provide the best picture for cross-city evaluation.

Furthermore, we applied our cross-city prediction model on the City of Rochester and analyzed qualitatively and visually what our models recommend and analyze why they suggest these vacant lot conversions.

Chapter 4

Results

This chapter discusses the results from the experiments described in the Section 3.4. For each of the experiments we analyze the classifiers built from the training set, and the discuss the results obtained from applying them on the validation set.

4.1 Baltimore Dataset

The training set for Baltimore consisted of a random sample of 60% of the data. The remaining 40% was used to validate our prediction model. The vacant lots in Baltimore had four available conversions. QCMOS represented lots that were converted to qualified community open spaces, ADOPTED represented lots that were converted to community gardens, URBAN FARM represented lots that were converted to urban farms and AVAILABLE represented vacant lots that were not converted.

We started by creating three separate simplified Random Forest classifiers for a subset of features to analyze how they interact with our target variable. As can be seen in Table 4.1, public utility such as distance from library, park etc. provides a strong indication of the vacant lot conversions, while

Feature	ADOPTED			Overall
	Precision	Recall	F1	Accuracy
publicUtil: libDist+parkDist+schoolDist	0.89	0.81	0.85	0.89
transitDist+category	0.66	0.62	0.64	0.75
vacantDensity+crimeDensity	0.80	0.75	0.77	0.84

Table 4.1: Performance of feature subsets for a Random Forest classifier on Baltimore

transit distance and zoning comparatively has weaker association. However, their prediction accuracy is still significantly better than random, and hence their contribution cannot be ignored.

mtry	Accuracy	Kappa	AccuracySD	KappaSD
2	0.81	0.68	0.01	0.02
8	0.83	0.72	0.01	0.02
14	0.83	0.72	0.02	0.03

Table 4.2: Hyperparameter tuning for Random Forest classifier on Baltimore

	Actual	Actual	Actual	Actual
Predictions	ADOPTED	AVAILABLE	QCMOS	URBAN FARM
ADOPTED	171	19	4	0
AVAILABLE	24	389	15	3
QCMOS	7	0	71	0
URBAN FARM	0	3	0	57

Table 4.3: Confusion matrix for the Baltimore dataset with Random Forest

4.1.1 Random Forest

The first experiment focused on using a Random Forest classifier to build our prediction model. We tuned our classifier using the number of trees in our Random Forest as a parameter $mtry$. As given in Table 4.2, the highest accuracy was provided with 8 decision trees in our Random Forest.

	Precision	Recall	F1	Balanced Accuracy
Class: ADOPTED	0.88	0.85	0.86	0.90
Class: AVAILABLE	0.90	0.95	0.92	0.91
Class: QCMOS	0.91	0.79	0.85	0.89
Class: U	0.95	0.95	0.95	0.97

Table 4.4: Statistics of predictions by target variable class for Random Forest on Baltimore

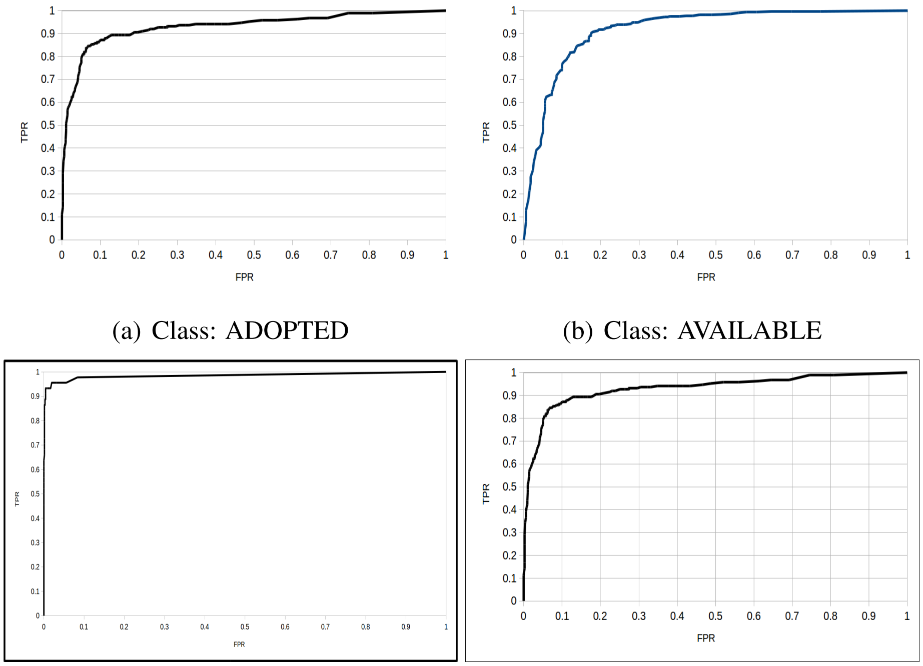
Figure 4.1 displays the ROC curves for our predictions for our validation set for Baltimore. Three of the classes (AVAILABLE, ADOPTED, QCMOS) have decent ratio of true positive rate to false positive rates. In the case of URBAN FARMS, our model performs exceptionally well on the validation set. A possible reason for this could be that urban farms have a more sophisticated plan in their development and much stronger zoning restrictions than the other classes of conversions. As shown in Table 4.3, the confusion matrix given also indicates similar results. Table 4.4 indicates good precision and recall for our system, and the balanced accuracy is approximately 90% for our classes.

size	Accuracy	Kappa	AccuracySD	KappaSD
3	0.59	0.29	0.04	0.04
5	0.63	0.40	0.04	0.06
8	0.66	0.42	0.03	0.04
10	0.69	0.47	0.01	0.02

Table 4.5: Hyperparameter tuning for MLP classifier on Baltimore

	ADOPTED	AVAILABLE	QCMOS	URBAN FARM
ADOPTED	148	72	22	0
AVAILABLE	42	315	28	1
QCMOS	11	13	40	0
URBAN FARM	1	11	0	59

Table 4.6: Confusion matrix for the Baltimore dataset with MLP



(a) Class: ADOPTED

(b) Class: AVAILABLE

(c) Class: URBAN FARM

(d) Class: QCMOS

Figure 4.1: ROC Curve for Random Forest classifier on the Baltimore dataset

	Precision	Recall	F1	Balanced Accuracy
Class: ADOPTED	0.61	0.73	0.67	0.78
Class: AVAILABLE	0.82	0.77	0.79	0.78
Class: QCMOS	0.62	0.44	0.52	0.70
Class: URBAN FARM	0.83	0.98	0.90	0.98

Table 4.7: Statistics of predictions by target variable class for MLP on Baltimore

4.1.2 MLP

The second experiment focused on the use of a Multilayer Perceptron (MLP) to build our prediction model. After performing hyperparameter tuning, the best results were obtained using an MLP with 10 hidden nodes as indicated by the variable $size$ in Table 4.5.

Table 4.6 gives us the confusion matrix for the classifier, while Table 4.7

gives us a class wise indication of the precision and recall of our prediction model. In terms of performance, MLP does not perform as well as a Random Forest. Its performance in classifying QCMOS is particularly inaccurate, and again URBAN FARM has a better classification accuracy than the other classes.

4.1.3 Naive Bayes

	ADOPTED	AVAILABLE	QCMOS	URBAN FARM
ADOPTED	102	5	5	0
AVAILABLE	97	399	49	3
QCMOS	3	0	36	0
URBAN FARM	0	7	0	57

Table 4.8: Confusion matrix for the Baltimore dataset with Naive Bayes

	Precision	Recall	F1	Balanced Accuracy
Class: ADOPTED	0.91	0.50	0.65	0.74
Class: AVAILABLE	0.73	0.97	0.83	0.77
Class: QCMOS	0.92	0.40	0.56	0.70
Class: URBAN FARM	0.89	0.95	0.92	0.97

Table 4.9: Statistics of predictions by target variable class for Naive Bayes on Baltimore

The third experiment focused on the use of a Naive Bayes classifier to build our prediction model. Table 4.8 gives us the confusion matrix for the classifier, while Table 4.9 gives us a class wise indication of the precision and recall of our prediction model. In terms of performance, Naive Bayes performs slightly better than MLP, but does not perform as well as a Random Forest. Its performance in classifying QCMOS is particularly inaccurate, and again URBAN FARM has a better classification accuracy than the other classes.

4.1.4 k-NN

	ADOPTED	AVAILABLE	QCMOS	URBAN FARM
ADOPTED	178	17	4	0
AVAILABLE	17	390	10	1
QCMOS	7	3	76	0
URBAN FARM	0	1	0	59

Table 4.10: Confusion matrix for the Baltimore dataset with k-NN

	Precision	Recall	F1	Balanced Accuracy
Class: ADOPTED	0.89	0.88	0.89	0.92
Class: AVAILABLE	0.93	0.95	0.94	0.93
Class: QCMOS	0.88	0.84	0.86	0.91
Class: URBAN FARM	0.98	0.98	0.98	0.99

Table 4.11: Statistics of predictions by target variable class for k-NN on Baltimore

The fourth experiment focused on the use of a k-Nearest Neighbor classifier to build our prediction model. The hyperparameter tuned for this particular classifier was the number of neighbors to consider for similarity of a vacant lot. The most optimal result was obtained for $k = 1$.

Table 4.10 gives us the confusion matrix for the classifier, while Table 4.11 gives us a class wise indication of the precision and recall of our prediction model. k-NN outperforms Random Forest for the Baltimore dataset in terms of precision, recall and accuracy. Due to the use of similarity between vacant lots k-NN does mimic the process an urban planner might take in choosing to convert a vacant lot, and as a result is better able to capture the pattern.

4.1.5 SVM

	ADOPTED	AVAILABLE	QCMOS	URBAN FARM
ADOPTED	106	33	20	0
AVAILABLE	86	364	31	1
QCMOS	9	5	39	0
URBAN FARM	1	9	0	59

Table 4.12: Confusion matrix for the Baltimore dataset with SVM

	Precision	Recall	F1	Balanced Accuracy
Class: ADOPTED	0.67	0.52	0.59	0.72
Class: AVAILABLE	0.76	0.89	0.82	0.78
Class: QCMOS	0.74	0.43	0.55	0.71
Class: URBAN FARM	0.86	0.98	0.91	0.98

Table 4.13: Statistics of predictions by target variable class for SVM on Baltimore

The fifth experiment focused on the use of an SVM classifier to build our prediction model. Table 4.12 gives us the confusion matrix for the classifier, while Table 4.13 gives us a class wise indication of the precision and recall of our prediction model. In terms of performance, SVM performs similar to MLP, but does not perform as well as a Random Forest and k-NN. Similar to other classifiers, it has a high accuracy for URBAN FARMS but suffers in terms of prediction accuracies for other classes.

The final results from our experiments are given in Table 4.14. As can be seen from the table, k-NN and Random Forest provides the best overall accuracy. Naive Bayes performs well only for specific classes but lacks in other areas. MLP and SVM only perform well for the URBAN FARM class prediction. In general, all classifiers have a significant accuracy when it comes to predicting URBAN FARMS, due to the systematic nature of such

Classifier	ADOPTED			AVAILABLE			QCMOS			URBAN FARM			Overall
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1	Accuracy
Random Forest	0.88	0.85	0.86	0.90	0.95	0.92	0.91	0.79	0.85	0.95	0.95	0.95	0.90
k-NN	0.89	**0.88**	**0.89**	**0.93**	0.95	**0.94**	0.88	**0.84**	**0.86**	**0.98**	**0.98**	**0.98**	**0.92**
SVM	0.67	0.52	0.59	0.76	0.89	0.82	0.74	0.43	0.55	0.86	**0.98**	0.91	0.74
MLP	0.61	0.73	0.67	0.82	0.77	0.79	0.62	0.44	0.52	0.83	**0.98**	0.90	0.74
Naive Bayes	**0.91**	0.50	0.65	0.73	**0.97**	0.83	**0.92**	0.40	0.56	0.89	0.95	0.92	0.78

Table 4.14: Summary of results for the Baltimore dataset

a conversion.

4.2 Philadelphia Dataset

Classifier	ADOPTED			Overall
	Precision	Recall	F1	Accuracy
Random Forest	**0.90**	**0.93**	**0.92**	**0.90**
k-NN	0.85	0.84	0.84	0.84
SVM	0.67	0.72	0.69	0.68
MLP	0.63	0.79	0.70	0.66
Naive Bayes	0.76	0.72	0.74	0.74

Table 4.15: Summary of results for the Philadelphia dataset

The dataset for Philadelphia was also split into a training & testing set consisting of a random sample of 60% of the data, while the remaining 40% were left for validation purposes. Each of the experiments that were performed on Baltimore dataset was performed again on the Philadelphia dataset.

The overall results are given in Table 4.15. For the Philadelphia dataset, there were only two classes as either vacant lots were AVAILABLE or they were ADOPTED. Similar to the results in Baltimore, Random Forest and k-NN performs the best in predicting the vacant lot conversions. This could be due to our use of categorical variables in normalization, as we convert all

Classifier	ADOPTED			Overall
	Precision	Recall	F1	Accuracy
Random Forest	**0.58**	0.19	0.28	0.48
k-NN	0.55	**0.40**	**0.46**	**0.49**

Table 4.16: Prediction statistics with Baltimore as training set

Classifier	ADOPTED			Overall
	Precision	Recall	F1	Accuracy
Random Forest	0.44	**0.53**	**0.48**	0.47
k-NN	**0.46**	0.51	**0.48**	**0.50**

Table 4.17: Prediction statistics with Philadelphia as training set

the data into categories based on the pre-defined utility scores.

4.3 Cross City Predictions

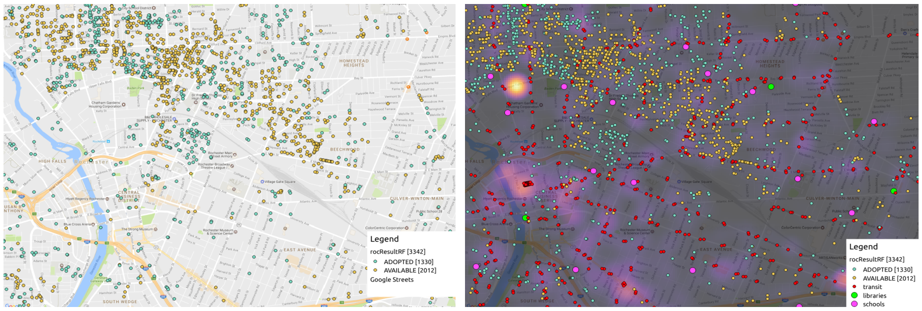
4.3.1 Baltimore and Philadelphia

For our cross city prediction model, we converted the dataset of Baltimore to point towards a binary class, indicating whether a vacant lot has been converted or not. The Philadelphia dataset already consists of binary classes, and as a result we performed two experiments. In the first one, we trained our model using data from Baltimore and performed predictions on Philadelphia. In the second one, we trained our model using data from Philadelphia and performed predictions on Baltimore.

The results of the experiments are given in Table 4.16 and Table 4.17. As can be seen, the precision, recall and accuracy are significantly low for cross-city predictions i.e. they are no better than random. One possible reason could be that each city has their own vacant lot programs and as a result patterns do not necessarily match up.

Another possible reason could be that the predictions made by our model could indicate vacant lots that have the indications of being a beneficial conversion, but simply has not been converted yet in another city. Further evaluation by expert urban planners and decision makers for our predictions would give a better indication.

4.3.2 Rochester



(a) Vacant Lot Predictions (b) Predictions with Features

Figure 4.2: Predictions by the Cross-City Classifiers

We utilized both Baltimore and Philadelphia dataset as our training sets, to create a Random Forest classifier to predict vacant lot conversions. We applied the classifier to vacant lot data for the City of Rochester. As can be seen in Figure 4.2(a), predictions were made as to which vacant lots should be converted. The green points indicate the lots that should be converted. Due to the lack of evaluation data and ground truth for the city of Rochester, we simply analyze why the predictions were made based on our feature set.

In Figure 4.2(b), all the other features are displayed including crime density in the background as a semi-transparent heatmap. The converted vacant lots are clustered towards higher crime rate zones, and follow along with the transit path. Furthermore, a lot of the clusters for the vacant lots that should be converted are farther rather than near to public utilities. This prediction goes against our conjecture in Section 3.1 of increased utility indicating increased likelihood of conversion. This result needs to be further evaluated by expert urban planners and validated with ground truth for further understanding of why these vacant lots were picked.

Chapter 5

Conclusions

5.1 Applications

The primary contribution of this work is to build and analyze a vacant lot model that can be used to predict future vacant lot conversions based on historical conversions of vacant lots. Our prediction models can be optimized further and integrated as a part of a larger urban planning system. The goal of our recommendation system is not to provide a complete solution but to be a part of a larger tool that would help support decision making for cities. Furthermore, our model can also be deployed as a part of a vacant lot toolkit, that would recommend to members of the community on vacant lots that may have a greater impact if they adopt or convert it.

5.2 Future Work

Our research can be further extended with the use of a greater number of cities, as we limited the scope of our research to only two city datasets. Furthermore, there is scope for improvement in our prediction model with the use of other non-stationary determinants such as satellite imagery, geographical weight etc. Our system can also be further evaluated by validating predictions made by expert urban planners, who can assess which vacant lots will have the most impact. In the future, it would also be an interesting project to attempt to build a generalized prediction model, instead of models specific to cities.

5.3 Conclusion

The objectives of this research were the development and design of a general dataset defining a vacant lot model that can be used for building recommendations or predictions for future vacant lot conversions, and development and evaluation of such a prediction model on two example datasets. The vacant lot model we built consisted of determinants such as distance to nearest public infrastructure, crime density, access through public transit, zoning policies etc.

We built our model for two example datasets, for the cities of Baltimore and Philadelphia, and built our prediction models for each of these cities on a portion of the datasets. We validated them against the remaining dataset, and found that our model captured vacant lot determinants and impact extensively well for cities with a more centralized approach to the vacant lot problem, while the accuracy was less for cities with more decentralized approaches. We also discovered that Random Forest and k-NN classifiers performed significantly better than other classifiers, due to their tendency to favor nominal variables.

Our prediction models displayed that it's feasible to have automatic recommendations as a starting point for tackling the vacant lot problem. Community leaders can use our model to pick a vacant lot to convert, while urban planners can use our system for a more macro level approach in terms of targeting specific vacant lots. The contributions of this thesis can be summarized as follows:

Vacant Lot Model We developed and designed a vacant lot model, that consists of features that determine if a vacant lot is converted. This model can be further extended with additional features of interest in the future.

Vacant Lot Conversions We utilized multiple classification models in existing literature to model vacant lot conversions as a data problem. We used the classification models to analyze if it's feasible to use historical vacant lot data, to recommend vacant lot conversions.

Bibliography

[1] BaltimoreHousing Adopt-A-Lot. http://www.baltimorehousing.org/adopt_a_lot. Accessed: 2017-04-21.

[2] Data.gov. https://www.data.gov/. Accessed: 2017-04-19.

[3] Grounded in Philly. http://www.groundedinphilly.org/. Accessed: 2017-04-21.

[4] Ramesh A., Anusha J., and Clarence J.M. Tauro. A novel, generalized recommender system for social media using the collaborative-filtering technique. *SIGSOFT Softw. Eng. Notes*, 39(3):1–4, June 2014.

[5] John Accordino and Gary T Johnson. Addressing the vacant and abandoned property problem. *Journal of Urban Affairs*, 22(3):301–315, 2000.

[6] Richard F Babcock, Charles L Siemon, et al. *The zoning game revisited*. Lincoln Institute of Land Policy Cambridge, MA, 1985.

[7] Charles C Branas, Rose A Cheney, John M MacDonald, Vicky W Tam, Tara D Jackson, and Thomas R Ten Have. A difference-in-differences analysis of health, safety, and greening vacant urban space. *American journal of epidemiology*, page kwr273, 2011.

[8] Grace W Bucchianeri, Kevin C Gillen, and Susan M Wachter. Valuing the conversion of urban greenspace. *pdf]. Available at:¡ http://phsonline.org/media/resources/Bucchianeri_Gillen_Wachter_Valuing_Conversion_Urban_Green KG_changesacceptes. pdf¿(accessed 27 November 2015)*, 2012.

[9] Joan Capdevila, Marta Arias, and Argimiro Arratia. Geosrs: A hybrid social recommender system for geolocated data. *Information Systems*, 57:111–128, 2016.

[10] Eugenia C Garvin, Carolyn C Cannuscio, and Charles C Branas. Greening vacant lots to reduce violent crime: a randomised controlled trial. *Injury Prevention*, 19(3):198–203, 2013.

[11] Ruining He, Chen Fang, Zhaowen Wang, and Julian McAuley. Vista: A visually, socially, and temporally-aware model for artistic recommendation. *arXiv preprint arXiv:1607.04373*, 2016.

[12] Megan Heckert and Jeremy Mennis. The economic impact of greening urban vacant land: a spatial difference-in-differences analysis. *Environment and Planning A*, 44(12):3010–3027, 2012.

[13] Mikhail Kanevski, Alexei Pozdnoukhov, and Vadim Timonin. Machine learning for spatial environmental data. *Theory, applications and software*, page 377, 2009.

[14] Peleg Kremer, Zoé A Hamstead, and Timon McPhearson. A social–ecological assessment of vacant lots in new york city. *Landscape and Urban Planning*, 120:218–233, 2013.

[15] Hillit Meidar-Alfi. Measuring the utility of urban infrastructure systems: A step towards a comprehensive evaluation of non-transportation infrastructure systems. 2009.

[16] Mohammad A Tayebi, Mohsen Jamali, Martin Ester, Uwe Glässer, and Richard Frank. Crimewalker: a recommendation model for suspect investigation. In *Proceedings of the fifth ACM conference on Recommender systems*, pages 173–180. ACM, 2011.

[17] Susan M Wachter and Kevin C Gillen. Public investment strategies: How they matter for neighborhoods in philadelphia. *unpublished report of the Wharton School of the University of Pennsylvania*, 2006.