https://doi.org/10.48009/1_iis_2020_55-61

Issues in Information Systems
Volume 21, Issue 1, pp. 55-61, 2020

USING MACHINE LEARNING ALGORITHMS TO ANALYZE IMPACT OF CRIME ON PROPERTY VALUES

Preslav Angelov, Seattle University, angelovp@seattleu.edu
Huy Le, Seattle University, lehuy@seattleu.edu
Evin Tolentino, Seattle University, tolenti4@seattleu.edu
Ben Kim, Seattle University, bkim@seattleu.edu

ABSTRACT

Since predictions of property values require an integration of a great deal of data, machine learning can play an important role. There are many factors influencing the sale price of private properties. In this paper, we focus on how different types of crimes affect the residential property sale price in an urban county in the USA - Pierce County, Washington. We are interested in what machine learning algorithms can be more effective in predicting the sales values. We worked on two sets of data. The first data source includes the physical attributes of a property, such as the square footage, quality, the year built and/or remodeled. The second data source contains the crime data for Pierce County from July 2018 to July 2019. We combined these two data sources into one full dataset. Then, we divided the full dataset into three clusters using the EM (Expectation Maximization) algorithm. The full and three clustered datasets were used to build two sets of data mining models – one with crime data, and the other without them. We built our models using three algorithms – decision trees, artificial neural networks, and random forests. In total, we created 32 models and evaluated them by calculating prediction errors of each model. We found that the random forest models produced the lowest values of errors. Using those models, we concluded that crime is a significant factor in predicting the sale price of residential properties in Pierce County, Washington.

Keywords: Property Values, Crime, Machine Learning, Random Forests, Clustering

INTRODUCTION

In this paper, we are interested in assessing how crimes affect the sales price of a residential value. Housing values are determined by the characteristics of a property as well as its surrounding environmental factors. One of those environmental factors is crime or safety of the neighborhood. For our research, we used the data available from a county located the Washington State in the USA (Pierce County). We worked on two sets of data. The first data source includes the physical attributes of a property, such as the square footage, quality, the year built and/or remodeled. The second data source contains the crime data of Pierce County from July 2018 to July 2019. For analyzing these data, we used three machine learning algorithms – decision trees, neural networks, and random forests.

Prediction and assessment of sales price of properties have many important implications. In the US, local property taxes are based on the assessed property values. From the government and taxpayers' perspective, accurate and fair estimations of the property values have direct impact on taxes to be levied and paid. Also, as we can see in the flourishing Internet-based real estate companies such as Zillow and Redfin, determination of the accurate property values is becoming a commercially indispensable competency for those companies. We believe that our research on predicting the property values using machine learning algorithms can make contributions for these types of applications. Fair and timely estimations of property values can help both buyers and sellers of properties in fast changing real estate markets.

LITERATURE REVIEW

Foryś and Putek-Szeląg (2017) discussed if there was a relationship between crime rate or the sense of security for a given area surrounding Szczecin, Poland, and housing prices. The authors indicated that the analysis of crime rates on housing prices also incorporated the elements of criminology and social sciences as well as statistical methods. Their goal was to gain a more comprehensive understanding of the relationship between crime rates and housing prices. The authors first analyzed the crime data provided by the Regional Police Headquarters in Szczecin. This information allowed the authors to understand the population of a region and the types of crime that occurred most in those areas. Moreover, the authors examined the property market of Szczecin to divide the regions in Szczecin, as defined by the Regional Police Headquarters. As a result, the authors were able to segment the descriptive statistics for housing prices, given the police station districts of Szczecin.

From this exploratory analysis, the authors were able to perform a pattern analysis to further investigate the crime rates in a given police station district and compare those findings against the changes in housing prices in Szczecin. As a result, the authors determined that areas with low-priced apartments connected with their location (transaction prices) coincide with the areas of rising non-pecuniary crimes. Additionally, the authors identified that a rise in housing prices was associated with subsequent increases in property crimes, which is mostly the result of the increased number of new households with higher incomes (the effect of gentrification). Overall, the authors concluded that an increase in housing prices for a given location corresponded with a decrease in the number crime, but simultaneously was susceptible to a higher number of property crimes.

Olijade and Lizam (2016) explained how crimes affected the property values of residential areas. They also discussed how the different types of crimes, such as burglary, street crimes, vandalism, affected the property values. The authors used the logistic regression analysis to predict the impact of the various forms of crimes on residential property values. The data used for developing the model were obtained from the multiple surveys collected from the residential neighborhoods within Southwestern Nigeria. Crimes were divided into the multiple independent variables: street crime, vandalism, robbery, and violent crime. All of the predictors showed a direct influence on property values; however, the greatest contributor negatively affecting the residential property values was violent crimes. The authors argued that violent crimes have the highest impact on property values because such crimes produced fear in residential communities. They concluded that the findings of their logit model were consistent with other published papers; that is, the model supported the hypothesis that residential neighborhood crimes negatively influence the residential property values. This negative relationship can deter housing investments, which in turn deteriorates the neighborhood and a reduction in property tax that can be used to fight residential crimes.

Yao and Fu (2017) used the data on houses and crimes that were collected from the Denver Open Data Catalog. For the housing dataset, the authors decided to use only the single-family detached homes. They selected 3000 houses evenly located in a major residential region of north Denver, including North Park Hill, South Park Hill, Hale, Montclair, and Hilltop. All the appraised values were recorded in 2015. For crime data, the authors collected the residential forcible burglaries that happened in Denver five years before the appraisals were conducted (2009-2014). The house profiles were collected from the demographic data of 2000 and 2010 of the US census survey. The authors discussed how to analyze the crime data to generate an in-depth understanding of community safety, or how the community safety of an area were assessed by analyzing historical crime data. They considered crime severity and crime temporal correlations. To relate the crimes to particular houses, they calculated the distance between the house and the location where the crimes were committed. The crimes were attributed to a house if they were committed within a specific range of distance from the house. To build the model, they adopted RankLib as baseline algorithm implementation. Also, they discussed how to model the impacts of community safety on house values without including the attributes such as the income levels and the ratings of nearby schools.

DATASET DESCRIPTION

We used two datasets located at https://www.co.pierce.wa.us/736/Data-Downloads and https://gisdata-piercecowa.opendata.arcgis.com/datasets/crime-data. These datasets were downloaded on July 8, 2019. Table 1 shows the database tables and columns for each table we used. One dataset involved the property values and tax information for Pierce County in the Washington state. This dataset contained the following tables: appraisal account, improvement, improvement built as, improvement details, land attribute, sale, seg merge, tax account, and tax information. The second dataset contained the crime information for Pierce County in the Washington state. This dataset contained the following information: CaseNo (A unique ID given to each crime case), City, District, LocCode (General description of the location of the crime), NAT_Name (Name of the Neighborhood Action Team (NAT) area where the crime occurred. A "None" value indicates that the crime did not occurred in a NAT area), OBJECTID

(Internal feature number), OccurredOn (Date of occurrence), Public_Nam (General description of the type of crime), XCoord (X-coordinates of location), YCoord (Y-coordinates of location), latitude, and longitude. Additionally, we included another dataset called Address Point that contained ZIP code information for Pierce County, Washington.

Table 1. Tables integrated and their respective columns

Appraisal Account	Improvement	Improvement Builtas	Tax Account	Address Point	Crime	Sale
ParcelNumber	BuildingID	PhysicalAge	AccountType	ZipCode	ObjectID	SalePrice
AppraisalAccountType	PropertyType	YearBuilt	TaxableValuePriorYear		OccurredOn	SaleDate
Buildings	SquareFeet	YearRemodeled	TaxableValueCurrentYear		Public_NAM	
LandGrossSquareFeet	PercentComplete				Latitude	
LandNetSquareFeet	Condition				Longitude	
AppraisalDate	Quality					
Latitude						

Table 2. New Crime Categories

Original Name	Category	Original Name	Category
Arson - Non-residential	Property	Motor Vehicle Theft	Property
Arson - Residential	Property	Possession of Stolen Property	Property
Assault - Aggravated	Personal	Robbery - Business	Property
Assault - Simple	Personal	Robbery - Other	Property
Burglary - Non-residential	Property	Robbery - Residential	Property
Burglary - Residential	Property	Robbery - Street	Property
Criminal Traffic	Other	Telephone Harassment	Personal
Drug Possession (Methamphetamine)	Drug	Theft - Gas Station Runout	Property
Drug Possession (Other)	Drug	Theft - Mail	Property
Drug Sale/Manufacture (Methamphetamine)	Drug	Theft - Other	Property
Drug Sale/Manufacture (Other)	Drug	Theft - Vehicle Prowl	Property
Fraud or Forgery	Property	Theft -Shoplifing	Property
Homicide	Homicide	Trafficking in Stolen Property	Property
Intimidation	Personal	Vandalism - Non-residential	Personal
Liquor Law Violations	Other	Vandalism - Residential	Personal
		Warrant Arrests	Other

DATA PREPROCESSING

Null Values

The null value ratios for the data columns were extremely low for those attributes we selected to use for our modeling. However, we took further steps to remove null values in the critical columns for data analysis. For example, any null values in the 'AppraisalAccountType' column are eliminated by filtering for only rows having ‘Residential’ values. The same principle was applied to the 'PropertyType' and 'AccountType' columns. Furthermore, we chose to focus our research only on the specific zip codes due to the availability of our crime data. This naturally took out any null values from our 'ZipCode'. These columns are shown in Table 1 as discussed earlier. We used SQL for all our operations to remove null values.

Data Integration

Table 1 shows the list of the tables and columns we chose to use. We used 'ParcelNumber' for joining the tables about the properties. However, crime data do not use a parcel number for identifying the rows of data; in fact, the crime dataset has no property identification numbers. Instead, it uses 'ObjectID', the latitude, and longitude to identify the criminal incident and its location. To integrate crime data into property data for evaluating community safety of properties, we used the distance between the location where crimes occurred and a specific property. If a crime occurred within one mile radius area from the property, this crime was counted in the crime attribute of the property. We believed that the number of committed crimes occurred around a given property could reflect the level of public safety more accurately than the crime rate of a wider area such as city or zip code. However, this method required too many calculations. We had approximately 20,000 rows for properties and 27,000 rows for crime. this would lead to approximately a half-billion calculations to obtain the entire data set of properties and crime in Pierce County. Due to the constraints of time and resources, we reduced our crime data by randomly selecting ten percent of the original crime dataset.

The purpose of this project is to understand the impact of crimes on property value in Pierce County. We used the most recent sale price to represent the current house value. We organized the various crimes into five categories: personal crime, property crime, drug crime, homicide, and other crime (see Table 2). Our final dataset included thirteen attributes for properties and five crime attributes as can be seen in Table 3.

Table 3. Final Dataset Attributes

Property Attributes	Crime Attributes
ParcelNumber	DrugCrime
Buildings	Homicide
LandNetSquareFeet	OtherCrime
SquareFeet	PersonalCrime
Condition	Other Crime
Quality	
PhysicalAge	
YearBuilt	
YearRemodeled	
TaxableValuePriorYear	
TaxableValueCurrentYear	
Sale Price	
Sale Date	

Correlation Analysis

As can be seen in Figure 1, we built a correlation matrix and identified highly correlated attributes. The results show that condition and quality have a high correlation coefficient. Among the attributes only 'PhysicalAge', 'YearBuilt' and 'YearRemodeled' are highly correlated with each other. Another interesting finding is that the types of crimes have high correlation coefficients among themselves. 'SalePrice' has negative correlation coefficients with all crime types.

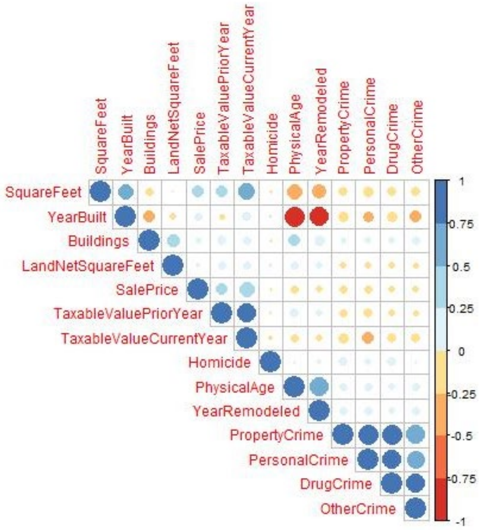

Figure 1. Correlation Matrix for the Variables used in Prediction Models

DATA MINING MODELS

We began our analysis using the two datasets, one with crime attributes and the other without crime attributes. With these two full datasets, we proceeded to use the following data mining models: decision trees, neural networks, and random forests, in order to explore whether crimes have an impact on property values in Pierce County, Washington. Before applying these algorithms, we divided the dataset into three clusters to make each cluster more homogeneous.

Clustering
Clustering can be used as a preprocessing method to enhance the homogeneity of the subset of data so that we can have a higher level of accuracy when applying machine learning algorithms. We used the EM (Expectation Maximization) algorithm to divide the original data into three clusters. Then, we applied the decision tree, neural network, and random forest algorithms on each cluster as well as the full dataset.

Decision Trees
Decision trees are one of the most popular algorithms used in business applications. They are built in a top-down recursive way by partitioning the data using the attributes providing the most information gains. The decision tree algorithm is rather popular in business applications because it can provide explanations about how a certain classification or numerical prediction was made by navigating the nodes on a decision tree.

We used two methods for splitting the data – binary and combined way. Binary splitting divides the data into two groups (e.g., high or low) while combined splitting allows multiple groups depending on the values of the splitting attribute. For choosing a splitting attribute, we used Bayesian Dirichlet Equivalent with Uniform prior as a scoring method as is provided by Microsoft Visual Studio. Figure 2 shows one of the decision trees we produced. We built sixteen decision trees in total.

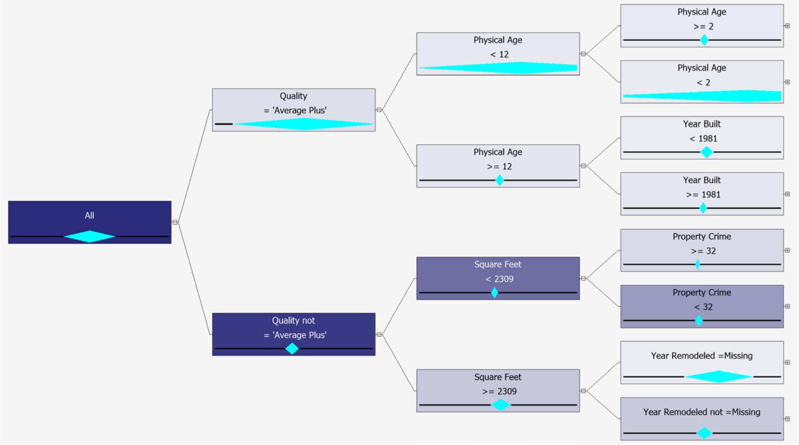

Figure 2. Decision Tree of Full Dataset Binary Split with Crime Attributes

Neural Networks
The Artificial Neural Network algorithm has been used to solve complex problems with non-linear relationships among the attributes for estimations or predictions. There appear to be no established heuristics on the optimal number of hidden layers or nodes on each hidden layer. We used one hidden layer and ten nodes on that hidden layer after some number of trials and errors. As explained in the next section, our neural network model did not produce as good prediction results as we initially expected. Figure 3 shows one of the neural networks we produced.

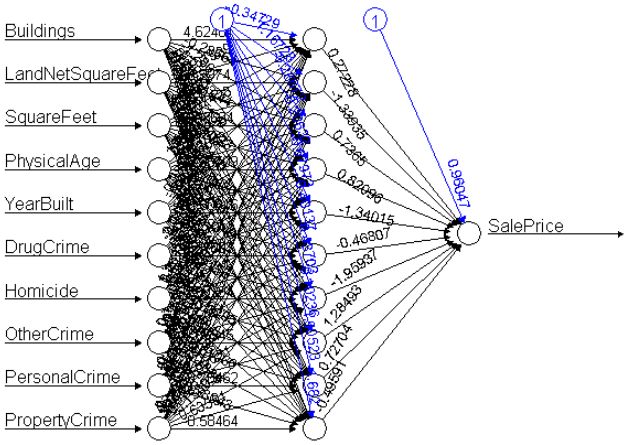

Figure 3. Neural Network with 1 Hidden Layer and 10 Nodes for Full Dataset with Crime Attributes

Random Forests
Instead of producing one decision tree, the random forest algorithm produces a multiple number of decision trees. Each tree is built by randomly selecting a subset of attributes. This random selection of attributes can reduce a prediction bias or overfitting by excluding attributes that may be highly correlated with each other. By calculating a mean of the predicted values of those decision trees, random forests produce a prediction. We used a default value of 500 for the number of decision trees to build a random forest model when using the 'randomForest' package in R. Using the random forest model, we are able to find out which attributes are more significant than others for predicting the sales price. As can be seen in the information gain plot of Figure 4, crimes were identified as the two most significant factors for determining the house value. The level of significance is calculated using the information gain as is used to build a decision tree. We produced the information gain plot for each cluster.

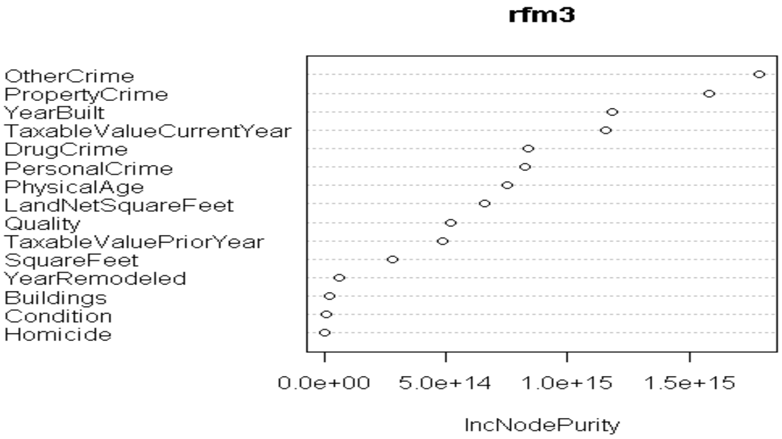

Figure 4. Information Gain for Random Forest Model from Cluster 3 with Crime Attributes

DISCUSSION AND IMPLICATIONS

For evaluations, we calculated Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE). We found that the random forest algorithm produced a smallest error. Also, the random forest models with the crime data produced a considerably lower errors than the ones without them as shown in Table 4.

However, this does not necessarily mean that the random forest model is the best algorithm or better than the two algorithms we used for property valuations. Depending on how to set different parameter values (e.g., the number of nodes on each hidden layer in the neural networks), we may have different values for errors.

As discussed earlier, we divided the data into three clusters. Among those clusters, cluster 3 had more criminal incidents than in other clusters. Using the random forest model, we performed an information gain analysis as shown in Figure 4. In cluster 3, 'Other Crime' and 'Property Crime' were two most significant factors in determining the sale price. Thus, in a crime-ridden area, we believe that a sale price cannot be accurately estimated or predicted without considering the crime data.

Table 4. RMSE & MAE Results from Random Forests

Random Forest	**With Crime**				**Without Crime**			
	Full	Cluster 1	Cluster 2	Cluster 3	Full	Cluster 1	Cluster 2	Cluster 3
RMSE	383,215	935,123	85,499	1,039,295	707,119	872,753	88,143	1,298,972
MAE	75,139	124,167	51,584	130,690	102,167	126,689	47,404	203,945
% Var Explained	68.64	21.45	78.41	67.41	49.92	16.52	57.88	55.74

We believe that our research has made the following contributions. First, the buyers of a property should be aware of the crime rate of the neighborhood where the property is located. Most of real estate companies or web sites do not necessarily show the crime rates. Most customers consider the physical attributes such as the size, number of bedrooms, and so on. However, buyers need to be aware of the crime rates before their purchase of the property. Second, sellers or neighbors needs to reduce the crime rate if they want to obtain the better price for their properties. The local police and other government agencies have an important role to play. Third, the Internet real estate agencies such as Zillow or Redfin should display the crime rates or display them more prominently so that prospective buyers can make appropriate and informed decisions.

CONCLUSIONS

In this paper, we used the physical characteristics of a property and crime data of its neighboring areas to assess or predict the value of a residential property using three machine learning algorithms – decision trees, neural networks, random forests - after dividing the original data into three clusters using the EM (Expectation Maximization) algorithm. We found that the random forest model produced the most accurate results for us. Also, using the information gains available from the random forest model, we identified the most to least significant attributes for predicting the property values.

Purchasing a house can be one of the most expensive investments in many people's life. Estimation of a fair market value is critical for buyers as well as sellers. We believe that machine learning algorithms can perform better than human assessors given the large amount of data provided. In addition to the physical attributes of a property, we included the crimes for our research. To produce more accurate estimations of property values, however, we believe that we should include more data such as the quality of schools, parks, hospitals, demographic data, and others. We chose three clusters, but more clusters could provide more accurate predictions. For the future projects, we plan to include more data as discussed above and experiment with a different set of parameter values such as various numbers of hidden layers for neural networks.

REFERENCES

Foryś, I & Putek-Szeląg, E., (2017). The Impact of Crime on Residential Property Value - on The Example of Szczecin. *Real Estate Management and Valuation*, 25(3), 51-61.

Olijade, S. E., & Lizam, M. (2016). Determining the Impact of Residential Neighborhood Crime on Housing Investment Using Logistic Regression. *Path of Science*, 2(12). Doi: 10.22178/pos. 17-13

Yao, Z., Fu, Y., Liu, B., & Xiong, H. (2016). The Impact of Community Safety on House Ranking. *Proceedings of the 2016 SIAM International Conference on Data Mining*. doi:10.1137/1.9781611974348.52