https://doi.org/10.48009/2_iis_2018_109-118
Issues in Information Systems
Volume 19, Issue 2, pp. 109-118, 2018

ANALYSIS AND PREDICTION OF REAL ESTATE PRICES: A CASE OF THE BOSTON HOUSING MARKET

Sharmila Muralidharan, Seattle University, muralidh@seattleu.edu
Katrina Phiri, Seattle University, phirik@seattleu.edu
Sonal K. Sinha, Seattle University, sinhas1@seattleu.edu
Ben Kim, Seattle University, bkim@seattleu.edu

ABSTRACT

In this paper, we estimate and predict the assessed prices of residential properties by applying machine learning algorithms (Decision Trees and Artificial Neural Networks) on the house properties and crime data. Property values have become an increasingly common topic of conversation in today's economy. After the financial crisis of 2007–2008 triggered by the United States housing market crash, the factors that determine housing prices have become of even greater interest to numerous parties, including government agencies, urban planners, developers, real estate professionals, finance professionals, and of course, most American homeowners. Overall housing market trends, the number of new home sales, and home-resales make up an important component of the U.S. economy. Consequently, data concerning these transactions is closely tracked for the purposes of determining economic activities and formulating appropriate monetary and fiscal policies. Companies such as Zillow, a leading real estate and rental marketplace provider, have grown in popularity as they focus on empowering consumers with data on the housing market. Zillow frequently releases housing market forecasts based on its database of more than 110 million U.S. homes - including homes for sale, homes for rent and homes not currently on the market, as well as "Zestimate home values", "Zestimates for Rentals" and other home-related information.

Keywords: Housing Market, Decision Trees, Neural Networks, and Linear Regression

INTRODUCTION

Positive factors currently affecting the US Housing Market are: moderately rising mortgage rates; low risk of a housing crash for most cities; millennial buyers coming into their main home buying years; a trend towards government deregulation; labor shortages pushing up costs of production and incomes; and the longest positive business cycle in history. The policies of the current administration are also expected to have an impact on housing, due to the focus on the repatriation of business, investment and jobs back to the US. Due to the high degree of fluctuation in the real estate market, investors, homeowners, appraisers, tax assessors and other real estate market participants, such as, mortgage lenders and insurers, are constantly speculating on housing prices. It is no wonder that there is a wealth of research on the topic.

Traditional house price prediction is based on cost and sale price comparisons. Forecasting is difficult as the factors that affect the housing market range from socioeconomic (e.g. per capita crime rate, access to transportation, average income, and educational attainment) to specific house features (e.g. square footage, number of bedrooms, building style, date of last remodel). Thus, the development and availability of various house price prediction models can play a helpful role in filling an information gap that can improve the efficiency of the real estate market. This paper attempts to build such models.

LITERATURE REVIEW

Bahia (2013) provides independent real estate market forecasts on home prices by using data mining techniques. The main idea was to construct the neural network model by using two types of neural networks. First feed forward neural network (FFBP) and second, Cascade forward neural network (CFBP). The two models are then compared to find the

best performing prediction of house prices. The authors estimated the median value of owner occupied homes in Boston suburbs, given 13 neighborhood attributes in a sample size of 506 data points.

Mu et al (2014) analyzed a dataset containing Boston suburb house values and use several machine learning methods to make forecasts. Based on these predictions, it is the aim of the authors that government agencies and real estate developers can make better decisions on whether to start real estate developments in corresponding regions. Home values are forecasted using the following methods: support vector machine (SVM), least squares support vector machine (LSSVM), and partial least squares (PLS) methods. These algorithms are compared to the predicted results. Using multiple characteristics, Boston housing values were forecasted and the it was found that the prediction results of the various machine learning approaches varied. Although serious nonlinearity exists within the data, the experiment results also showed that the SVM and LSSVM methods were superior to the PLS method in dealing with the problem of nonlinearity.

Hromada (2015) describes an innovative software that can be used for real estate evaluations and analysis of real estate advertisements published on the Internet in the Czech Republic. The software systematically collects, analyzes and assesses data about the changes in the real estate market. For each half year, the software assembles over 650,000 price quotations concerning sale or rental of apartments, houses, business properties and building lots. All real estate advertisements are continuously stored in a software database and are thoroughly analyzed for their credibility.

Jaen (2002) presents insights gained from predicting real estate property values by applying data mining algorithms such as Decision Trees and Neural Networks. The Multiple Listing System (MLS) database was used to obtain the relevant data which consisted of 1229 transactions involving real estate properties for the 1999 to 2001 period in the city of Coral Gables, Florida. The dataset was examined for outliers. As a result of the outlier analysis, properties below $100K and those above $700K were removed from the analysis. This reduced the dataset to only 959 cases and included properties with values ranging from $108K to $700K. The paper explains how the Neural Network algorithm works and the methods employed to make the neural network learn from the training dataset. The backpropagation algorithm is a common method of training artificial neural networks. A decision tree modeling technique, primarily CART algorithm, was selected because the dependent variable is continuous. Linear Regression analysis was also conducted first and the results indicated that the overall model significantly predicted the sales price.

BOSTON HOUSING MARKET

2016 was a great year for the Boston real estate market. Home prices and appreciation rates continued to exceed the national average. Despite this fact, the Boston housing market is still one of the most affordable in the nation, with homeowners paying far less than the national average. Figure 1 shows the distribution of property types in the data set

to be described below. We see that 38% of the properties in Boston are residential condo units, 19% are single family residential properties, 11% are 2-family residential and other property types constitute 32%.

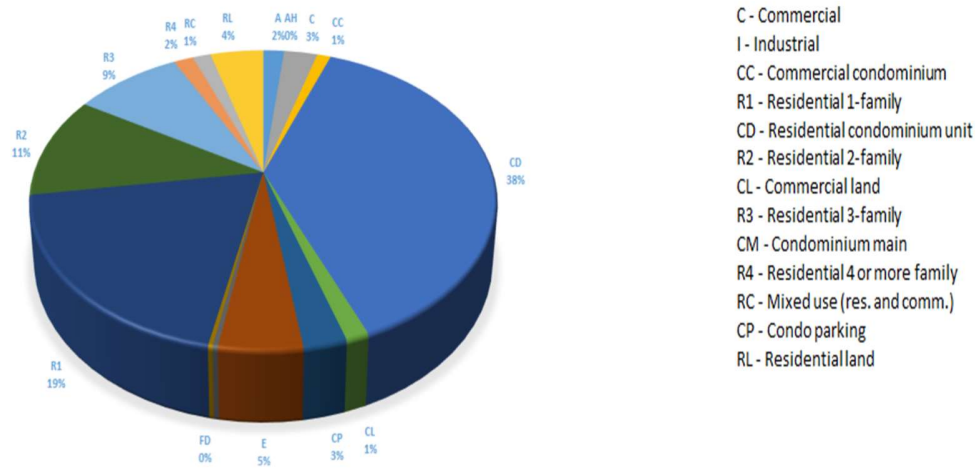

Figure 1. Property Types in Boston Dataset (2016 data from City of Boston)

Below are some quick facts about the Boston market (for Q1, 2016)

- Median Home Price: $378,500
- 1-Year Appreciation Rate: 1.0%
- 3-Year Appreciation Rate: 13.9%
- Unemployment Rate: 4.0%
- 1-Year Job Growth Rate: 1.9%
- Population: 667,137
- Median Household Income: $75,667

According to Trulia, an online residential real estate site for home buyers, sellers, renters and real estate professionals, the average Boston home was worth approximately $422/sqft in Q1 of 2016. This represented a decrease of eight percent from the same period in 2015. Furthermore, the median home price for Boston real estate was $378,500 during Q1, 2016, while the national average was $215,767. Additionally, appreciation rates for the Boston market also performed better than the national average, with home values appreciating 13.9% over the prior 12 quarters. Below is a comprehensive breakdown of the Boston market appreciation (Merrill, 2016):

- Homes purchased in the Boston housing market one year ago appreciated, on average, by $10,187 (national average - $15,781, over the same period).
- Homes purchased in the Boston housing market three years ago appreciated, on average, by $64,281 (national average - $49,356, over the same period).
- Homes purchased in the Boston housing market five years ago appreciated, on average, by $81,064 (national average - $68,727, over the same period).
- Homes purchased in the Boston housing market seven years ago appreciated, on average, by $129,577 (national average - $59,758, over the same period).
- Homes purchased in the Boston housing market nine years ago appreciated, on average, by $40,273 (national average - $16,435, over the same period).

DESCRIPTION OF DATASETS

The data was gathered from the website of the City of Boston (http://cityofboston.gov). We have chosen two datasets – assessed property prices and crime statistics, and correlated the two using geolocation data from Google Web Services. The property assessment data is for the year 2016, and the crime data includes recent history of crime in the City of Boston for the period 2012-2015. We have a total of 169K data points for the property assessments and the data set contains several property types and detailed characteristics of each property that amount to 76 predictors and the outcome variable which is the assessed property price.

The crime data set contains 268K data points, consisting of 20 variables, including the crime type and geolocation information. About the dependent variable, the assessed price for the property includes the assessed price for the land and the building which are also separately available in the dataset. To answer our first problem statement, we concentrated only on the assessed property prices dataset. We partitioned the dataset into residential (single-family homes & other) and non-residential (condominium units) based on the property type and included the variables that were relevant to the corresponding property types. Figure 2 shows how we processed the datasets from collection to model validation.

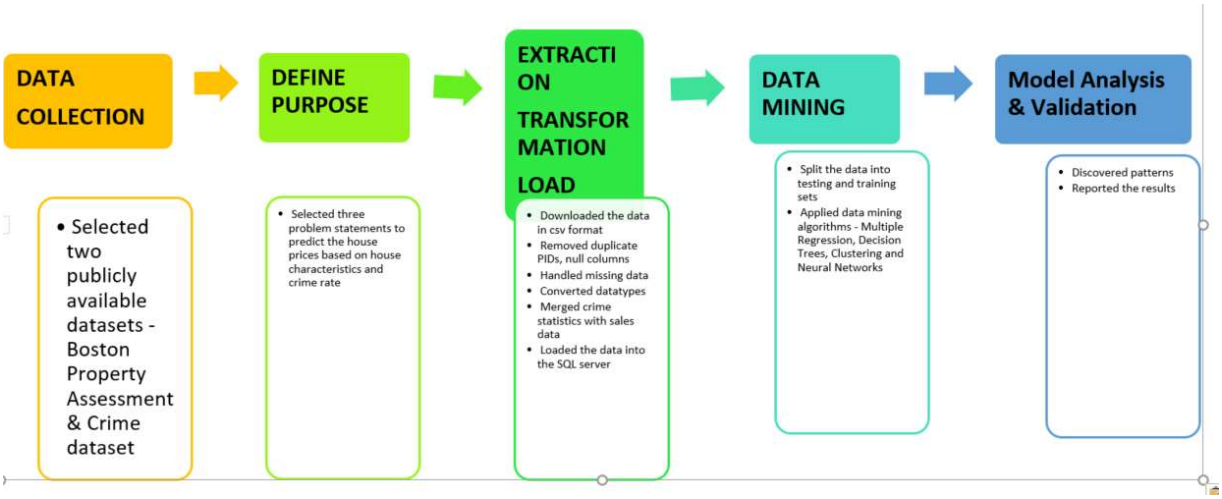

Figure 2. Stages for Data Mining to Build Housing Price Prediction Model

DATA PREPROCESSING

Data Reduction

The size of the initial dataset was 44 MB and contained approximately 169,000 rows. We performed initial data pre-processing in Microsoft Excel before importing the data into SQL Server Management Studio. First, the null ratio for each column was calculated for all 77 columns of the Boston Property Assessment 2016 data in MS Excel and columns with high null values were deleted. The attributes mentioned below were eliminated from our dataset as they were not relevant to any of the three problem statements above and/or they had high null values:

Outlier Analysis

This was conducted on all the numeric columns in R Studio on the reduced Boston property dataset after the above operations. Both the box plot analysis and the quantile analysis were run in R Studio to find the outliers and then using these values, the acceptable range for each attribute was found. Then, the rows containing these outliers were identified in SQL by using the "w*here"* function for each attribute and the related tuples were deleted. For example, attributes such as YR_BUILT & YR_REMODL were limited between 1900 and 2016 (Refer Appendix for detailed code on outlier analysis).

CrimeRateComputed	Zipcode	TotalCrime	TotalPopulation	Neighborhood
0.684476534	2108	2844	4155	Beacon Hill/Financial District
0.003290051	2467	75	22796	Chestnut Hill
0.431722689	2111	3288	7616	Chinatown/Financial District/Leather District
0.34015748	2210	648	1905	Fort Point
0.234310539	2125	7497	31996	Dorchester-Uphams Corner-Savin Hill
0.120744155	2131	3667	30370	Roslindale
0.174528763	2134	3574	20478	Allston
0.222551988	2124	11098	49867	Dorchester-Codman Square-Ashmont
0.173887539	2114	2075	11933	Beacon Hill/West End
0.058914729	2199	76	1290	Prudential Center
0.329202712	2109	1408	4277	North End
0.109941039	2135	4270	38839	Brighton
0.409126542	2119	9916	24237	Roxbury
0.149496196	2115	4362	29178	Back Bay/Fenway-Kenmore
0.144968332	2129	2472	17052	Charlestown
0.136589206	2136	3991	29219	Hyde Park
0.263159964	2122	6694	25437	Dorchester-Fields Corner
0.008686211	2163	16	1842	Allston-Harvard Business School
0.090858889	2132	2468	27163	West Roxbury
0.901076716	2110	1339	1486	Financial District
0.296613018	2118	7943	26779	South End
0.14668906	2128	6114	41680	East Boston
0.141376878	2130	5212	36866	Jamaica Plain
0.176974836	2127	5760	32547	South Boston
0.259654491	2121	6959	26801	Dorchester-Mount Bowdoin
0.104590434	2113	761	7276	North End
0.42147481	2116	8985	21318	Back Bay/Bay Village
0.116488891	2215	2763	23719	Fenway-Kenmore
0.245517137	2120	3245	13217	Mission Hill

Figure 3. Crime Rate for Each Neighborhood

Data Integration

In the crime table, we only had geometry information and hence reverse geocoding was done to obtain the zip codes for each crime location. After obtaining the zip code, two new columns were added to the Crime dataset namely, Neighborhood and Population details based on the zip code obtained from reverse geocoding. Neighborhood data was obtained from Boston Census information and population data was obtained from MassAnalysis (https://statisticalatlas.com/place/Massachusetts/Boston/Population). Neighborhood data was added based on the zip codes. Data Integration of the Boston Property Assessment table and the Crime table was done by doing an *INNER JOIN* on the zip code column. Figure 3 shows the computed crime rate for each zip code. Hence, the initial Boston property table had 3 derived columns namely Neighborhood, Population and Crime Rate Computed (see Figure 4).

Based on the correlation matrix for numeric variables and domain knowledge of the categorical variables, we were able to select features for further model building. Correlation test function was used both in R and JMP to create a correlation matrix for numeric variables.

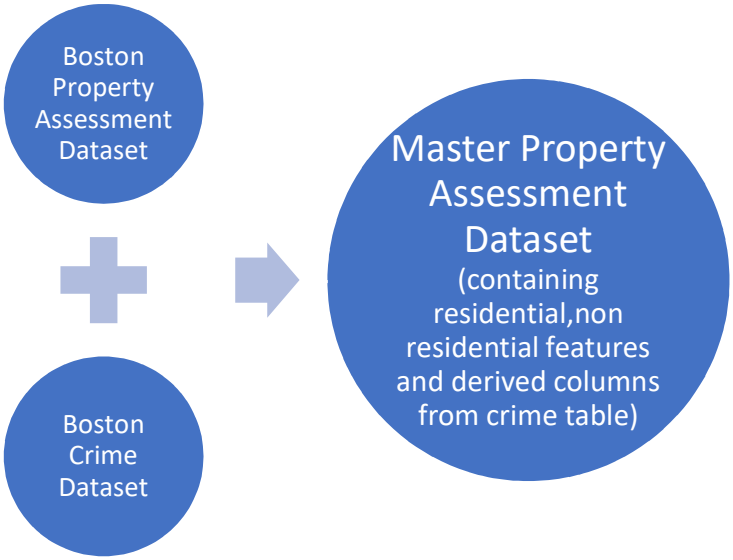

Figure 4. Building the Final Datasets

Finally, for all tables, numerical data was transformed using min-max normalization in SQL and problem-statement specific views were created for model building and evaluation.

ANALYSIS

In selecting the attributes to include in the mining models, we chose ones that would produce the most significant results for prediction and classification. After various experimentation in MS Visual Studio, all attributes were carefully considered after analyzing their impact on the data mining models. Eventually, 13 variables were selected for data mining as shown in Figure 5 below.

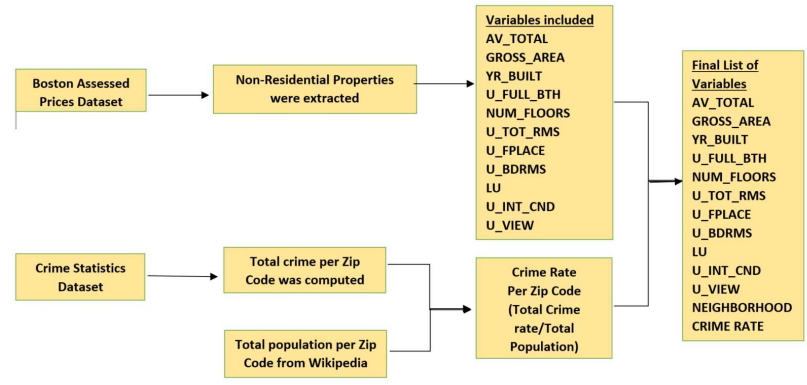

Figure 5. Attributes used for models

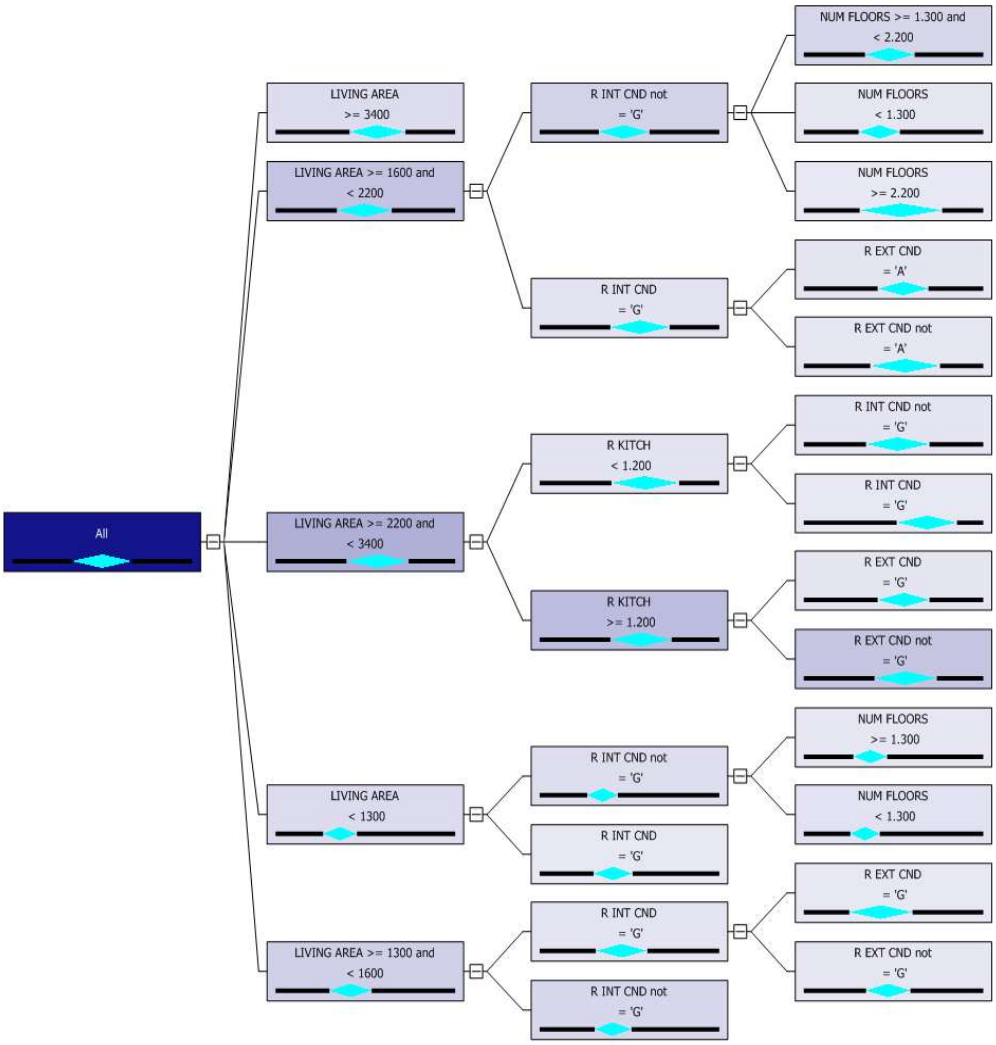

Figure 6. Decision Trees Mining Model Viewer

Decision Trees

The Decision Tree Viewer for AV_TOTAL shows the decision rules that were created in the model. Going from "ALL", which is the root of the decision tree model, we can see the various decision nodes of the tree. The boxes that are darker shades of blue represent the variables that have a stronger impact on the assessed value of a residential property. In this case, it is clear that level 1, living area (LIVING_AREA), has the greatest effect on assessed value. From there, the model branches off to the level 2 variables, which are the next most influential – in this case, interior condition "G-good" (R_INT_CND) and kitchen (R_KITCH). Interestingly, we see that kitchen is only significant for one node. The level 3 variables are the number of floors (NUM_FLOORS), the exterior condition "G-good" & "A-average" (R_EXT_CND), and interior condition is significant once again. The Dependency Network for assessed value shows that the strongest dependency links are living area, full bath (R_FULL_BTH), fireplace (R_FPLACE), and number of floors.

Linear Regression

In calculating the assessed value of a residential property in Boston, the linear regression equation was as follows:

AV TOTAL = 425,260.504+34,616.960*(R FULL BTH-1.832)+24,350.772*(NUM FLOORS-1.946)-13,222.593*(R KITCH-1.623)-10,502.476*(R BDRMS-4.380)+13,684.292*(R HALF BTH-0.372)+40,933.963*(R FPLACE-0.334)+64.155*(LIVING AREA-2,193.072)

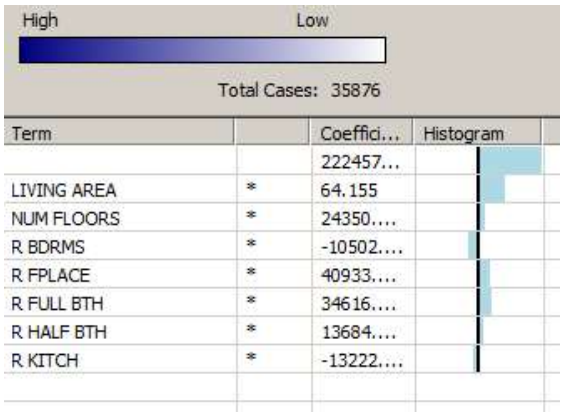

Figure 7. Linear Regression Coefficients

Neural Networks

With the Neural Network model, we were able to determine which variables and their respective probability contribute the most to assessed values of residential homes. The neural network model discretizes the assessed values into 4 ranges, which for our purposes, we can call low (0.000-0.343), average (0.343-0.465), above average (0.465-0.586), and high (0.586-1.000). Arranging the factors in descending order of probability of having either low or high assessed value, we found the following:

- 74.49% probability that houses with interior condition "excellent" will have high assessed value
- 63.86% probability that homes with "good" exterior finish will have high assessed value
- 40.24% probability that houses with "poor" overall condition will have low assessed value
- 45.85% probability that houses with "Victorian" building style will have a high assessed value
- Surprisingly, the model also showed, with a 90.49% probability, that homes with a "luxury" bath style were likely to have low assessed value.

VALIDATION OF MODELS

In building our prediction models, it was important to validate the accuracy of each of our models. In order to determine whether our predictions were accurate, we measured various model error rates. Each of our models would be highly optimized for the data they were trained on, thus the expected error rate on testing data will be higher than that of the data on which it was trained. For our validation, we calculated three measures of error (see Table 1):

- MSD - Mean Squared Deviation
- MAD -Mean Absolute Deviation
- MAPE -Mean Absolute Percentage Error

After calculating MSD, MAD, & MAPE for both residential and non-residential datasets, the tables below were generated to show the output. In the case of both the residential and non-residential models, the linear regression model showed the lowest error rate. However, it must be noted that due to the fact that linear regression cannot deal with categorical variables, this is not an entirely accurate picture. Between the decision trees and neural network models, neural networks performed better.

Table 1. Measures of Errors

Models	Residential (single-family) House Features without Crime Rate		
	MSD	**MAPE**	**MAD**
NN (Neural Network)	0.02776587	30.6454	0.1274102
DT (Decision Tree)	1.38E+10	22.84374	94203.47
MLR (Multiple Linear Regression)	1.36E+10	22.1721	91912.95

CONCLUSION

From our exploratory data, technical statistical interpretations and data mining predictions, we can make some conclusions about our data. For example, there is no statistical significance between year remodeled (YR_REMODL) and kitchen style (R_KITCHN_STYL) with total assessed value (AV_TOTAL). However, we found a statistical significance between the crime and housing price. From the multiple linear regression, we obtained the equation for problem statement three, PRICE = 451,552.323*(Crime Rate-0.208) as a model for predicting price given the crime rate. Realtors have practical experience that enables them to assume that if home or a condominium unit is located on a corner lot, this factor will have a statistical significance on the assessed value of that property. These same results can be used to determine what other factors influence assessed value. For example, on average a house with excellent interior condition and view will be worth more than one with an average view and condition. Based on the predicted values, an investor or realtor has more information to make a decision to sell a house or to wait until a certain value is reached. All of these results are important aspects that should be considered when building or selling a home.

Future work on this research would involve conducting more tests and enhancements on the models and the attributes. Firstly, we expect to improve accuracy by possibly doing further transformation of variables to decrease the error rates. Secondly, we would also like to download the shape files of Boston, convert the Boston map into grids and overlay the crime and average house price data points on the map so that each grid shows the number of crimes and the average house price information. Further, we would like to obtain geographic and income information from the Census Bureau for the City of Boston and see how these factors influence property prices. Additionally, if we are provided access to historical sales data of these parcel ids, we would be able to better forecast the behavior of the housing market in the long-term. Granted, this would entail using larger sets of historical sales data to capture the cyclical nature of the housing market.

REFERENCES

Bahia, I. S. (2013). A Data Mining Model by Using ANN for Predicting Real Estate Market: Comparative Study. *International Journal of Intelligence Science*, 162-169. Retrieved April 2017, from http://file.scirp.org/pdf/IJIS_2013100815382055.pdf

Hromada, E. (2015). Mapping of Real Estate Prices Using Data Mining Techniques. *Procedia Engineering*, 233-240. Retrieved April 2017, from https://www.researchgate.net/publication/283334152_Mapping_of_Real_Estate_Prices_Using_Data_Mining_Techniques

Jaen, R. D. (2002). Data Mining: An Empirical Application in Real Estate Valuation. American Association for *Artificial Intelligence*, 314-17. Retrieved April 2017, from http://www.aaai.org/Papers/FLAIRS/2002/FLAIRS02-062.pdf

Merrill, T. (2016, Sept 21). Boston Real Estate Market & Trends. Retrieved June 4, 2017, from Fortune Builders: http://www.fortunebuilders.com/boston-real-estate-market-trends/

Mu, J., Wu, F., & Zhang, A. (2014). Housing Value Forecasting Based on Machine Learning Methods. Abstract and Applied Analysis, Volume 2014 (2014), Article ID 648047, 7 pages. Retrieved April 2017, from https://www.hindawi.com/journals/aaa/2014/648047/