Issues in Information Systems
Volume 16, Issue IV, pp. 30-36, 2015

DATA MINING ANALYSIS AND PREDICTIONS OF REAL ESTATE PRICES

Victor Gan, Seattle University, gany@seattleu.edu
Vaishali Agarwal, Seattle University, agarwal1@seattleu.edu
Ben Kim, Seattle University, bkim@Taseattleu.edu

ABSTRACT

In this paper, we analyzed the real estate transaction data, and built prediction models for the real estate price using data mining algorithms, and validate the models. The datasets we used were publicly available from King County in the State of Washington - the datasets were for the year 2012 and 2013. Each has 22,812 and 27,333 transactions, respectively. We built two models using two algorithms - decision trees and neural networks - and compared their performances.

Keywords: Data Mining, Real Estate Data Analysis, Decision Trees, Neural Networks

INTRODUCTION

Starting from year 2008, home values in King County suffered a steep fall in the sales price followed by a strong recovery as shown in Figure 1. Per the data from Zillow [12] through January 2015, the median home price in King County is $415,200, almost recovering back to the peak median home price of $434,000 reached in July/August 2007, just before the recession hit. From the low median home price of $311,000 from September 2011 to January 2012, the price has steadily increased to $415,200 in January 2015. This represents a significant 33.5% over a time period of 3 years increase from the low we observed in January 2012 to the current median home price.

The King County residential real estate market presents an interesting opportunity for us to analyze and predict where home prices are moving towards. Zillow has predicted a 5.3% increase for the next year from Jan 2015 to Jan 2016.

Zillow uses market data and statistics to come up with a method to estimate home values. "The Zestimate® home valuation is Zillow's estimated market value, computed using a proprietary formula." "The *Zestimate* is automatically computed three times per week based on millions of public and user-submitted data points" [13]. Based on what was quoted from the Northwest Multiple Listing Service [11], median home prices sold in King County increased by 14% from 2012 to 2013.

With the availability of publicly recorded real estate sales transactions from the King County Assessor's Office, we built the models to predict home prices using the Decision Trees and Neural Networks algorithms. We used the transaction data from 2012 to create mining models and then predicted the year 2013 prices using a random sample of 1,000 homes. Calculating the differences between the actual and predicted price as an error, we compared the results and accuracy of the predicated prices by the Decision Trees and Neural Networks algorithms.

The real estate market in the United States has shown some striking trends in the last couple of years. Prediction of real estate prices is becoming increasingly important. Real estate prices not only reflect the overall market conditions but are also a good indicator of economic health. Data mining techniques are gaining an importance in the field of predictive modeling. Our datasets obtained from the King County website had 82 variables, 47 of which we used after removing obviously irrelevant variables.

LITERATURE REVIEW

There were many attempts to predict real estate prices using different methods. Khashan [5] used primarily the linear regression and text mining algorithms to predict the real estate prices in Dubai. Using text mining in addition to regression analysis, they were able to reduce prediction errors.

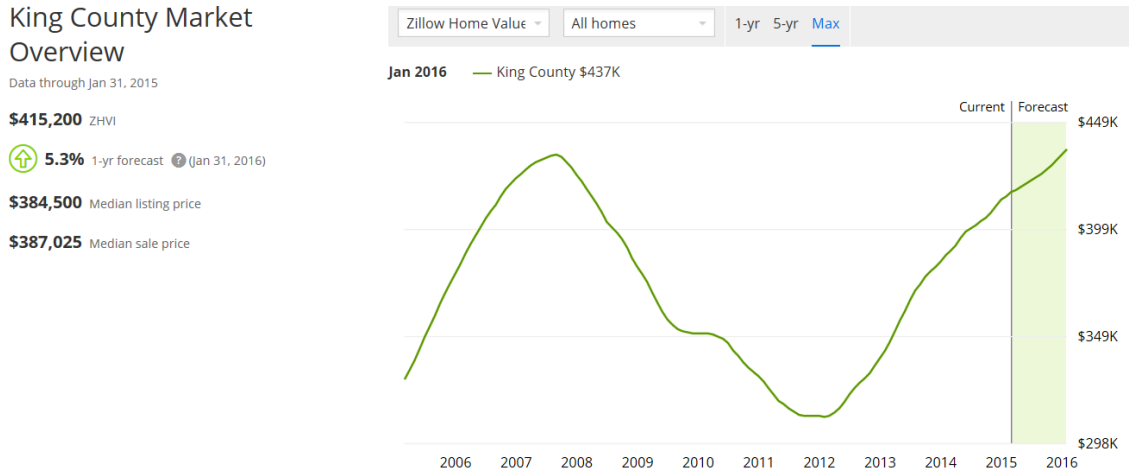

Figure 1. Price Trend of Home Values in the King County Market Per Zillow [13]

Acciani, Fucilli, & Sardaro [1] analyzed real estate prices and they concluded that the Model Tree and Multivariate Adaptive Regression Splines performed well even with small datasets, and underlined the fact that data mining modeling techniques deserved further research to obtain in-depth knowledge and understanding of real estate markets.

Del Cacho [2] compared several algorithms for the problem of housing valuation. He used both hedonic and non-hedonic pricing models. He argued that a group of model tree algorithms (un-bagged and bagged) works well for mass appraisals in urban areas and produced more precise predictions than the linear regression and neural network models. The K-nearest neighbors approach emerges to be the second best approach in his study. But ultimately it was the model tree bagged algorithm that provided the best results with 15% deviation from the quoted price.

Jaen (2002) examined the factors that determined the housing prices, and used knowledge discovery techniques such as neural nets and decision trees. He examined the dataset on real estate transactions to identify the factors that influence selling price, and proposed the development of a model that could be used to predict real estate prices. Additionally the writer states that the decision tree algorithm known as C&RT produced the best results and used the least number of predictors to arrive at the solution.

Ng et al [10] explored whether clustering methods have a role in spatial data mining. To evaluate the effectiveness of the data mining algorithm, they applied it to real estate data sets. For this they used their own cluster analysis algorithm called CLARANS. They found CLARANS to be more efficient than the clustering methods prevalent at that time.

Grether and Mieszkowski [3] identified the most important determinants of real estate values. They found that many structural characteristics, such as one story vs two story, or one-car garage vs two-car garage were important. Realtor's evaluation of the house was also significant. On the other hand there were many variables that were insignificant such as basement, insulation and dollar value of improvements.

EXTRACTION, TRANSFORMATION, AND LOADING (ETL)

We started the project by looking for data sources that were publicly available. While Zillow has real estate data sources available [14], it only contains data for one month, with the latest one being January 2015. The King County Assessor's office has real estate sales and transaction data available from

Table 1. Factors used for Analysis

	Format	Length
Major	character	6
Minor	character	4
Building Number	number	3
Number Living Units	number	1
Situs Address	character	
Building Number	Character	5

Fraction	Character	3
Direction Prefix	Character	2
Street Name	Character	25
Street Type	Character	4
Direction Suffix	Character	2
Zip code	Character	10
Stories	number	3
Building Grade	number	2
Building Grade Variation	number	2
Square Feet 1st Floor	number	5
Square Feet Half Floor	number	5
Square Feet 2nd Floor	number	5
Square Feet Upper Floor	number	5
Square Feet Unfinished Full	number	5
Square Feet Unfinished Half	number	5
Square Feet Total Living	number	5
Square Feet Total Basement	number	5
Square Feet Finished Basement	number	5
Finished Basement Grade	number	2
Square Feet Garage Basement	number	5
Square Feet Garage Attached	number	5
Daylight Basement	character	1
Square Feet Open Porch	number	5
Square Feet Enclosed Porch	number	5
Square Feet Deck	number	5
Heat System	number	2
Heat Source	number	2
Percent Brick Stone	number	3
View Utilization	character	1
Bedrooms	number	2
Bath: Half Count	number	2
Bath: 3qtr Count	number	2
Bath: Full Count	number	2
Fireplace: Single Story	number	2
Fireplace: Multiple Story	number	2
Fireplace: Freestanding	number	2
Fireplace: Additional	number	2
Year Built	number	4
Year Renovated	number	4
Percent Complete	number	3
Obsolescence	number	3
Percent Net Condition	number	3
Condition	number	1
Additional Cost	number	6

1951 to the most recent transaction date as of December 31, 2014. With such comprehensive availability, the dataset from King County was chosen for our data mining project.

Major sources of data were from the King County Recorder's Office and the Department of Assessments' Office. The King County Recorder's Office keeps records and provides access to over 350 types of documents, including official records of real estate sales transactions in King County. We have also used data downloaded from Zillow [12] to help us understand and analyze the real estate data. For King County, we downloaded all the real estate transaction data from the following website: http://info.kingcounty.gov/assessor/DataDownload/default.aspx

As shown in Table 1, a residential home real estate in the King County records are identified by the "Major" and "Minor fields." Major is a 6-digit numerical character field, and minor is a 4-digit numerical character field. Together, the Major and Minor fields are combined to form a unique 10-digit numerical field called [ParcelID], which is used by King County to uniquely identify every piece of real estate property. In total, there were 22,812 transactions in 2012 and 27,333 transactions in 2013, excluding the homes with sale price equal to 0 and no known zip codes.

We included only residential homes, and did not include any commercial real estate transactions for offices or retail shops. Within residential real estate transactions, we have only included homes that King County defined as

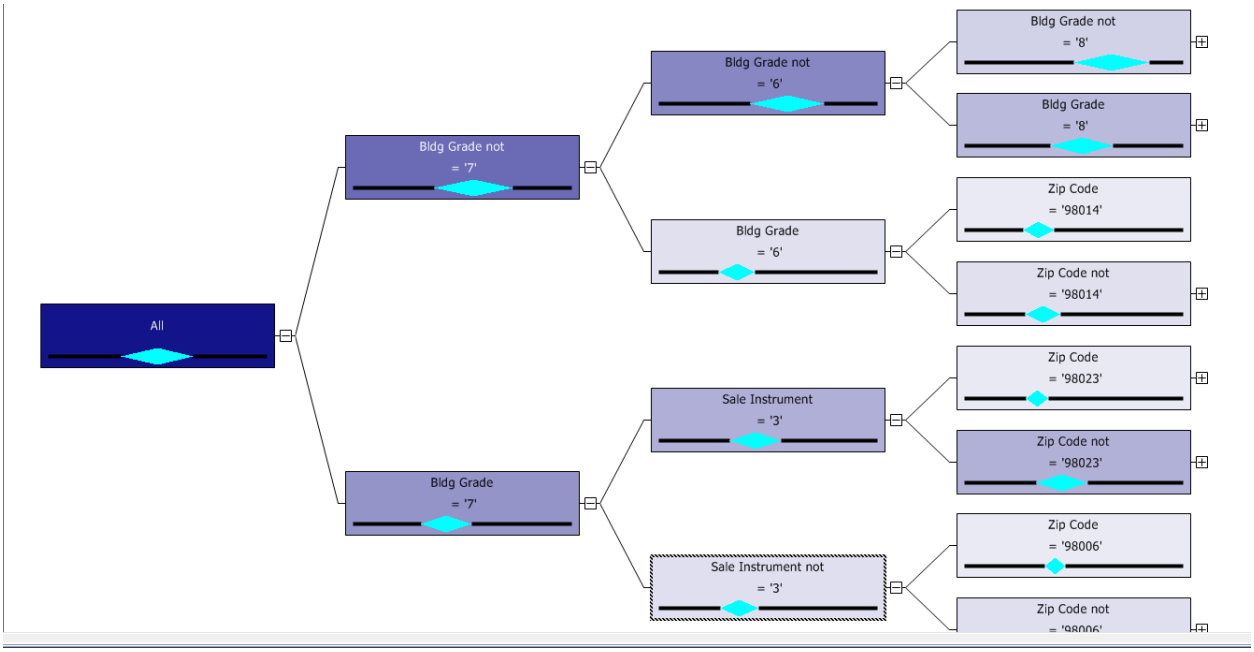

Figure 2. Decision Tree

residential buildings. Residential buildings are classified by King County as buildings with 1, 2 or 3 living units, and tend to be single-family homes. Apartments and condominiums are not included in this category.

For the transactions in year 2012 and 2013, we excluded the bottom 5% of the home sale prices, as well as the top 5% of the home sale prices. We found that the bottom 5% of the home sales prices, which were around $100,000 or below, tend to be outliers which were not representative of the home sale prices. Similarly the top 5% of home sale prices, which generally were greater than $1 million, tend to be outliers that skewed the accuracy of the model. For testing and validation, we generated 1,000 random samples from the 2013 transaction data.

TOOLS AND ALGORITHMS USED

For our analysis, we used Microsoft SQL Server 2014 – Database and Analysis Services to apply the Decision Trees and Neural Networks algorithm to the datasets described earlier.

Decision Trees

The Microsoft Decision Trees algorithm is a classification and regression algorithm provided by Microsoft SQL Server Analysis Services for use in predictive modeling of both discrete and continuous attributes. To build a decision tree, we used *entropy* to calculate the information gains to build a hierarchy of the decision tree. For discrete attributes, the algorithm makes predictions based on the relationships between the input columns in a dataset. It uses the values, known as states, of those columns to predict the states of a column that are designated as predictable. Specifically, the algorithm identifies the input columns that are correlated with the predictable column. The decision tree makes predictions based on this tendency toward a particular outcome. For continuous attributes, the algorithm uses linear regression to determine where a decision tree splits [6, 7].

Table 2. Comparison of Prediction Error Results between Microsoft Decision Trees and Neural Networks after Adjusting for 14% Increase in Sales Price

Predication Error after Adjustment	Microsoft Decision Trees	Neural Networks
Mean Absolute Error (MAE)	102,968	83,579
Standard Deviation	108,474	73,595

Table 3. Price Categories

Home Value (HV)	Category
HV < 200000	1
HV >= 200000 and HV < 300000	2
HV >= 300000 and HV < 400000	3
HV >= 400000 and HV < 500000	4
HV >= 500000 and HV < 600000	5
HV >= 600000 and HV < 700000	6
HV >= 700000 and HV < 800000	7
HV >= 800000 and HV < 900000	8
HV >= 900000 and HV < 1000000	9
HV >= 1000000	10

Neural Networks

The Microsoft Neural Network algorithm creates a network that is composed of up to three layers of neurons. These layers are: an input layer, an optional hidden layer, and an output layer. In the input layer, input neurons define all the input attribute values for the data mining model, and their probabilities. Neurons in the hidden layer receive inputs from input neurons and provide outputs to output neurons. The hidden layer is where the various probabilities of the inputs are assigned weights. A weight describes the relevance or importance of a particular input to the hidden neuron. The greater the weight that is assigned to an input, the more important the value of that input is. Weights can be negative, which means that the input can inhibit, rather than favor, a specific result. In the output layer, output neurons represent predictable attribute values for the data mining model [8, 9].

ANALYSIS AND VALIDATION

In our study we have built Decision Trees and Neural Networks models using the actual residential property transactions in King County for 2012. As described earlier, we used Microsoft Visual Studio SQL Server's Analysis Services to access the data mining database available on Microsoft SQL Server's Analysis Server. The transaction data were available on Microsoft SQL Server. Using the "Mining Model Prediction" tool available on Visual Studio, we produced the predictions for property transaction prices of year 2013. These prediction values were saved in SQL Server database to calculate the prediction errors from the decision trees and neural networks, respectively.

A decision tree produced is shown in Figure 2. Using this model we have predicted the price of a random sample of 1,000 actual residential property transactions in King County for 2013, and compared the predicted values with actual ones. From year 2012 to 2013, the median home price of King County homes went up by 14% as aforementioned. Since we are using 2012 sales transaction data to create our mining models, this means that the prices the models predicted will need to be adjusted upwards by 14% in order to match the overall increase.

In our analysis, we considered only the attributes of the property. Any economic conditions external to the properties were not factored in the analysis, e.g., the inflation rate, mortgage interest rate, disposable income, unemployment rate, etc.

Table 4. Neural Networks - Error Results by Price Category

Category	Mean Absolute Error	Standard Deviation
2	60,757	51,179.69
1	63,929	68,859.04
4	66,164	56,209.37
3	66,309	50,020.36
5	93,616	62,902.71
6	113,299	76,174.59
7	179,283	100,247.74

8	237,330	94,992.07
10	277,967	65,456.90
9	288,547	104,177.44

Table 2 shows the prediction errors measured in MAE (Mean Absolute Error) and standard deviation. We also wanted to find out which categories of home prices our Decision Trees and Neural Network models can predict more accurately. Subsequently, we categorized the home prices into 10 categories as shown in Table 3.

As summarized in Table 4, we found that the Neural Network model generates MAEs below the overall mean of $83,579 for Categories 2, 1, 4 and 3. This means that for homes that are below $500,000 (Categories 1 to 4), we were able to produce more accurate results than for homes with higher transaction prices.

As summarized in Table 5, the result for the Decision Trees is similar, with Categories 2, 3, 1 and 4 producing MAEs that are below the overall MAE of $102,967. Category 2, for homes in the $200,000 and less than $300,000 range, generates the lowest MAE for Decision Trees and Neural Networks. However, we can also see that overall, for each and every category, Neural Networks has a lower MAE than Decision Trees. We were able to predict home prices in the $200,000 to $300,000 range most accurately, and the worst category was category 9.

While both Decision Trees and Neural Networks produce the best results in Category 2, the order is slightly different after that. For Decision Trees, in order of accuracy, the order is Category 3, 1 and 4. For Neural Networks, the order is Category 1, 4 and 3. From Category 5 onwards, the order is the same, starting with Category 5, 6, 7, 8, 10 then 9. The Neural Networks algorithm is better at predicting home prices in the $100,000 to below $200,000 range, followed by $400,000 to less than $500,000, then $300,000 to less than $400,000. The Decision Trees algorithm, however, was better at predicting the $300,000 to less than $400,000 range of home prices first, followed by the $100,000 to less than $200,000 range, and then the $400,000 to less than $500,000 range. For higher home prices in the $500,000 and above range, the order is the same for both algorithms.

Based on the results, the Neural Networks algorithm generates a lower mean prediction error when compared to the Decision Trees. The results we obtained using the two algorithms (Decision Trees and Neural Networks) on our dataset was different from the findings of Jaen [4] and del Cacho [2], who found that the Decision Trees algorithm worked better than Neural Networks for real estate valuation. There could be a number of reasons for this, including different real estate data sets being used, different time periods, as well as different regions and locations. While we used King County's actual sales transactions, del Cacho used the data from the city of Madrid in Spain, and Jaen's findings were based on data from the multiple listing system (MLS) database, in the three years from 1999 to 2001.

CONCLUSIONS

We performed the ETL (Extraction, Transformation, and Loading) process on the data from 2012 King Country real estate transactions, and build the prediction models using Neural Networks and Decision Trees. Between the two algorithms, we found that the Neural Networks algorithm produced lower mean errors than Decision Trees.

Table 5. Decision Tree - Error Results by Price Category

Category	Mean Absolute Error	Standard Deviation
2	68,368	80,204.97
3	77,667	65,311.8
1	82,242	104,272.2
4	90,687	70,243.31
5	112,618	89,386.89
6	129,239	101,164.9
7	204,457	134,261.5
8	264,068	138,741
10	337,608	36,828.48
9	343,380	133,321.9

As the next step, we recommend using the 2013 transaction data to predict and compare the accuracy of the predicted prices with the actual prices for 1,000 random samples from 2014. Doing this will help validate whether the Neural Networks algorithm is indeed better for predicting King County home prices than Microsoft Decision Trees. We can also compare the accuracy of other data mining algorithms to predict home prices using the data set that is available from King County. In addition, we can use the same methodology we have adopted, to compare and contrast the results if we apply the algorithms data from other regions. Another direction of research can be to include the other relevant factors such as SAT scores, average income, crime rate, and others for each zip code for future analyses.

REFERENCES

1. Acciani, C., Fucilli, V., & Sardaro, R. (2011). Data mining in real estate appraisal: a model tree and multivariate adaptive regression spline approach. Aestimum, 0, 27-45.
2. del Cacho, C. (2010). A comparison of data mining methods for mass real estate appraisal.
3. Grether, D. M., & Mieszkowski, P. (1974). Determinants of real estate values. Journal of Urban Economics, 1(2), 127-145.
4. Jaen, R. D. (2002). Data Mining: An Empirical Application in Real Estate Valuation. In FLAIRS Conference (pp. 314-317).
5. Khashan, D. Y. A. (2014). Using Data Mining and Text Mining Techniques in Predicting the Price of Real Estate Properties in Dubai.
6. Microsoft Decision Trees Algorithm Technical Reference, https://msdn.microsoft.com/en-us/library/cc645868.aspx
7. Microsoft Decision Trees Algorithm, https://msdn.microsoft.com/en-us/library/ms175312.aspx
8. Microsoft Neural Network Algorithm Technical Reference, https://msdn.microsoft.com/en-us/library/cc645901.aspx
9. Microsoft Neural Network Algorithm, https://msdn.microsoft.com/en-us/library/ms174941.aspx
10. Ng, R. T., & Han, J. (1994, September). Efficient and Effective Clustering Methods for Spatial Data Mining. In Proc. of (pp. 144-155).
11. Northwest Multiple Listing Service, (2015). Retrieved May 11, 2015 from http://www.northwestmls.com/library/content/statistics/Recaps.pdf
12. Zillow King County (2015), http://www.zillow.com/king-county-wa/home-values
13. Zillow Zestimate (2015), http://www.zillow.com/zestimate
14. Zillow Research Data (2015), http://www.zillow.com/research/data