Expert Systems with Applications 42 (2015) 2928–2934

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Using machine learning algorithms for housing price prediction: The case of Fairfax County, Virginia housing data

Byeonghwa Park[a], Jae Kwon Bae[b,*]

[a] Department of Business Administration, Keimyung University, 1095 Dalgubeoldaero, Dalseo-gu, Daegu 704-701, Republic of Korea
[b] Department of Management Information Systems, Keimyung University, 1095 Dalgubeoldaero, Dalseo-gu, Daegu 704-701, Republic of Korea

ARTICLE INFO

Article history:
Available online 26 November 2014

Keywords:
Housing price index
Housing price prediction model
Machine learning algorithms
C4.5
RIPPER
Naïve Bayesian
AdaBoost

ABSTRACT

House sales are determined based on the Standard & Poor's Case-Shiller home price indices and the housing price index of the Office of Federal Housing Enterprise Oversight (OFHEO). These reflect the trends of the US housing market. In addition to these housing price indices, the development of a housing price prediction model can greatly assist in the prediction of future housing prices and the establishment of real estate policies. This study uses machine learning algorithms as a research methodology to develop a housing price prediction model. To improve the accuracy of housing price prediction, this paper analyzes the housing data of 5359 townhouses in Fairfax County, Virginia, gathered by the Multiple Listing Service (MLS) of the Metropolitan Regional Information Systems (MRIS). We develop a housing price prediction model based on machine learning algorithms such as C4.5, RIPPER, Naïve Bayesian, and AdaBoost and compare their classification accuracy performance. We then propose an improved housing price prediction model to assist a house seller or a real estate agent make better informed decisions based on house price valuation. The experiments demonstrate that the RIPPER algorithm, based on accuracy, consistently outperforms the other models in the performance of housing price prediction.

1. Introduction

The continuous rise in interest rates since 2005 has markedly slowed the housing market in the US Lehman Brothers Holdings, Inc., a US investment bank, was forced into bankruptcy on September 15, 2008 because of excessive borrowing of financial instruments that were devalued because of a serious reduction in housing prices. The insolvency of Lehman Brothers Holdings, Inc. and the sub-prime mortgage crisis intensified the slowdown of the actual economy and the decline of asset values. These depreciated the global real estate market and housing prices and sparked a global financial crisis.

House sales are determined based on the Standard & Poor's Case-Shiller home price indices and the housing price index of the Office of Federal Housing Enterprise Oversight (OFHEO). These reflect the trends of the US housing market. According to the Case-Shiller home price indices, housing prices have declined by approximately 30–60% in major cities in the US since the sub-prime mortgage crisis. In Los Angeles, home prices peaked in September 2006 and continued to fall until the end of 2011. In June 2011, the housing prices in Los Angeles fell to 61.46% compared to September 2006 when the prices were the highest. Similarly, in New York, housing prices peaked in June 2006 and continued to fall until the end of 2011. In June 2011, home prices in New York fell to 29.56% compared to June 2006 when the prices were the highest.

Beginning in November 2012, the US housing market is experiencing a rapid recovery because of the decreasing inventory of houses, the increasing demand for new houses following employment growth, and government policies supporting the real estate market. To sustain the recovery trend of the housing market, timely real estate policies from the government are required. Therefore, housing price indices determining the housing market trend must be researched and developed. Housing price indices can be important indicators for stakeholders in the real estate market including real estate agents, appraisers, assessors, mortgage lenders, brokers, property developers, investors and fund managers, and policy makers, as well as to actual and potential homeowners. Moreover, the development of a housing price prediction model would greatly assist in the prediction of future housing prices and the establishment of real estate policies. This study uses machine learning algorithms as a research methodology to develop a housing price prediction model.

Machine learning has been used in disciplines such as business, computer engineering, industrial engineering, bioinformatics, medical, pharmaceuticals, physics, and statistics to gather

[*] Corresponding author. Tel.: +82 53 580 6410; fax: +82 53 580 6364.
E-mail addresses: bhpark@kmu.ac.kr (B. Park), jkbae99@kmu.ac.kr (J.K. Bae).

http://dx.doi.org/10.1016/j.eswa.2014.11.040
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

knowledge and predict future events. With the recent growth in the real estate market, machine learning can play an important role to predict the price of a property. However, few researchers have experimented on the selling price for real estate properties using machine learning algorithms. In the real estate market, real estate agents, buyers, and sellers are all important players. If homeowners wish to sell their townhouse, they can be represented by a real estate agent. The agent inputs the information regarding the seller's townhouse into a Multiple Listing Service (MLS). Other real estate agents can then access this information as an active listing. Sellers desire to sell their townhouses at their asking price. Conversely, buyers attempt to pay less than the listing price to close the transaction. Therefore, there can be price differences between the listing price that the seller originally expects and closing price that the buyers pay. From a seller's point of view, if the closing price is greater than or equal to the listing price, the deal is profitable. If the closing price is less than the listing price, then it could be a loss for the seller. We use machine learning algorithms as a tool to predict whether the closing price will be greater or less than the listing price.

It is a well-known fact that housing price valuation is one of most important trading decisions affecting a national real estate policy. In this study, we create models using machine learning algorithms such as C4.5, RIPPER (Repeated Incremental Pruning to Produce Error Reduction), Naïve Bayesian, and AdaBoost (Adaptive Boosting) to predict housing price.

The remainder of this paper is organized as follows. Section 2 reviews some background research studies on housing price prediction. Section 3 explains experimental design and analysis procedure. Experimental results are presented and analyzed in Section 4, and finally, our concluding remarks are provided in Section 5.

2. Housing price prediction models

An analysis of the housing market and housing price valuation literature indicates two principal research trends: the use of the hedonic-based regression approach and artificial intelligence techniques for developing housing price prediction models. For several decades, various hedonic-based methods were utilized to identify the relationship between house prices and housing characteristics (Adair, Berry, & McGreal, 1996; Selim, 2009). Meese and Wallace (2003) developed hedonic-based regression approaches to evaluate the effect of market fundamentals on housing price dynamics. Stevenson (2004) re-examined heteroscedasticity in hedonic house price models in his study using the average ages of houses in Boston, Massachusetts as data. The results obtained supported the evidence of heteroscedasticity regarding the house age in the previous findings. Bin (2004) estimated a hedonic price function using a semi-parametric regression and compared the price prediction performance with conventional parametric models. The results revealed that semi-parametric regression provided improved performance in both in-sample and out-of-sample price predictions and that it could be used for the measurement and prediction of house prices (Bin, 2004). Kestens, Theriault, and Rosier (2006) investigated household-level data in hedonic models to measure the heterogeneity of implicit prices regarding household type, age, educational attainment, income, and the previous tenure status of the buyers, either first-time owner or former owner.

However, hedonic-based methods have potential limitations relating to fundamental model assumptions and estimation. These are the identification of supply and demand, market disequilibrium, the selection of independent variables, the choice of the functional form of hedonic equation, and market segmentation (Fan, Ong, & Koh, 2006; Schulz & Werwatz, 2004; Selim, 2009).

Recent studies have focused on price prediction performance comparison between hedonic-based methods and machine learning algorithms. Kauko, Hooimeijer, and Hakfoort (2002) examined neural network modeling with an application to the housing market in Helsinki, Finland. Their results indicated that various dimensions of housing sub-market formation could be identified by uncovering patterns in the dataset. Furthermore, they demonstrated the classification abilities of two neural network techniques: self-organizing map and learning vector quantization. Fan et al. (2006) suggested various tree-based approaches that provide an important statistical pattern recognition tool in examining the relationship between house prices and housing characteristics. Liu, Zhang, and Wu (2006) proposed a fuzzy neural network prediction model based on hedonic price theory to estimate the appropriate price level for new real estate. The experimental results indicated that the fuzzy neural network prediction model had strong function approximation ability and was suitable for real estate price prediction. Selim (2009) compared the prediction performance between the hedonic regression and artificial neural network models. This study demonstrated that artificial neural network models can be an improved alternative for prediction of house prices in Turkey. In another study, Kuşan, Aytekin, and Özdemir (2010) suggested a fuzzy logic model for prediction of the selling price of house-building. Azadeh, Ziaei, and Moghaddam (2012) introduced a hybrid algorithm based on fuzzy linear regression and a fuzzy cognitive map to address the problem of forecasting and optimization of housing market fluctuations. Over the past few years some researchers have studied on performance comparisons among machine learning algorithms to identify better forecasting models. Gerek (2014) suggested two different adaptive neuro-fuzzy (ANFIS) approaches for the estimation of house selling price in the construction sector. The experimental results indicated that the ANFIS with grid partition models performed better than the ANFIS with sub clustering models. The ANFIS with grid partition technique can be successfully used in the estimation of house prices in the construction sector. Wang, Wen, Zhang, and Wang (2014) suggested real estate price forecasting models based on particle swarm optimization (PSO) and support vector machine (SVM). The experimental results indicated that the proposed PSO–SVM based real estate price forecasting model has good forecasting performance compared to grid and genetic algorithms. Gu, Zhu, and Jiang (2011) suggested a hybrid of genetic algorithm and support vector machines (G-SVM) approach to forecast housing price. The experimental results indicated that forecasting accuracy of G-SVM is more superior to traditional methods. However, little research has been carried out on developing a better housing price prediction model through performance evaluations of several machine learning algorithms. This study aims at comparing performance of machine learning algorithms and developing more accurate housing price prediction model for the real estate market.

3. Experimental design

This section describes how to establish the experiment in order to test performance of machine learning algorithms for classification. We started from merging real estate, public school ratings, and mortgage rate data into an integrated dataset. Four machine learning classifiers were selected and tested on *WEKA* data mining software. To determine the performance of each classifier, we explored two performance tests which are three-way data split with 10-folds and 10-fold cross-validation.

3.1. Data source and selection

For the experiment, we used the real estate datasets from three different sources: MLS, historical mortgage rates, and public school ratings. Real estate data was obtained from the Metropolitan

Regional Information Systems (MRIS) database. This database is used nationwide by real estate agents in the US to advertise properties for sale. This database is a MLS and is updated by the individual agents and brokers.

MLS provides in-depth information regarding the different types of real estate markets. The focus of this paper is on the residential market, specifically townhouses in the County of Fairfax, Virginia. There are several types of listings in the MLS such as active, pending, withdrawn, and sold. We selected only records with "sold" status to ensure information with both the listing and closing price. We extracted 15,135 records from three different sources (i.e., MLS, historical mortgage rates, and public school ratings) spanning the period from January 1, 2004 to May 28, 2007. Each extracted record included 76 attributes (variables). Of the 76 variables, 49 were then selected using a *t*-test as a preliminary screening. Subsequently, 28 variables were selected by a stepwise logistic regression. Tables 1 and 2 summarize the variables and their definitions used in this study. Table 1 presents 16 variables describing the physical features of townhouses.

In Table 2, the variables "ElementarySchoolRate", "MiddleSchoolRate", and "HighSchoolRate" are associated with public school ratings; these data were collected from the "greatschools" website (www.greatschools.org). The primary goal of this website is to rate public schools and to provide information about every public school in the nation. This website provides a ranking for public schools that ranges from zero to ten, where ten is the best. We retrieved the rankings for all the public schools in Fairfax County.

Fixed mortgage rates (FMR) and adjusted mortgage rates (AMR), which are among the mortgage contract rate variables in Table 2, refer to the weekly mortgage rates. This data contains 15 and 30 year FMRs as well as 1 year AMRs. This data was retrieved from the "Mortgage-X Mortgage Information Service" website (www.mortgage-x.com).

3.2. Data cleaning and integration

Because the property data is entered manually by the individual real estate agents, there was missing and incorrect information. Some of the missing values were correctable using other attributes. For example, a missing city name or zip code was completed based on the address, looking up the correct values from the Fairfax County website. In a few records the listing price and/or closing price for the townhouse was listed as less than $5000. In these cases, we interpreted this value as an error and excluded the record. There were also records with incorrect units listed; we converted these records into the correct unit. For example, 2240 acres should have been 2240 square feet. Other examples of errors include the total square feet of the townhouse being smaller than the lot size, or the closing date being earlier than the listing date. In these cases, the associated record was removed.

Table 1
List of physical features variables selected.

Category	Name of attributes	Descriptions	Original data type
Physical features (16)	BasementTypeValue	Type of basement (fully finished; partially finished; full; partial; walkout)	Nominal
	Bathsfull	Number of bathroom having a toilet, wash basin and bathing facilities	Ratio
	Bathshalf	Number of bathroom having a toilet and wash basin	Ratio
	Bedrooms	Number of bedrooms	Ratio
	ExteriorTypeValue	Type of exterior (brick; aluminum siding; vinyl siding; wood/cedar; composition; brick front; stone; stucco; concrete)	Nominal
	ExteriorFeaturesTypeValue	Type of exterior features (deck; bump; fenced-fully; fenced-partial; fenced-rear)	Nominal
	CoolingTypeValue	Type of cooling system (central a/c; heat pump; ceiling fan; attic fan)	Nominal
	Fireplaces	Number of fireplaces	Ratio
	TotalSquare	Square feet of living area	Ratio
	GarageSpaces	The number of cars parked inside garage	Ratio
	HeatingTypeValue	Type of heating system (baseboard; electric air filter; forced air; forced air; heat pump; radiator)	Nominal
	HeatingFuelTypeValue	Type of heating fuel (bottled gas/propane; central; electric; natural gas)	Nominal
	HotWaterTypeValue	Type of fuel for hot water (bottled gas/propane; electric; natural gas)	Nominal
	StyleTypeValue	Style of the property (colonial; contemporary; split foyer; split level; etc.)	Nominal
	LotSqft	Square feet of lot size	Ratio
	ParkingType	Type of parking (garage; covered parking; driveway; unassigned; assigned street; street)	Nominal

Table 2
List of public school rating and mortgage rate variables selected.

Category	Name of attributes	Descriptions	Original data type
Public school ratings (3)	ElementarySchoolRate	Quality of elementary school where the property is located (1–10; 10 is the highest)	Ordinal
	MiddleSchoolRate	Quality of middle school where the property is located (1–10; 10 is the highest)	Ordinal
	HighSchoolRate	Quality of high school where the property is located (1–10; 10 is the highest)	Ordinal
Mortgage contract rate and others (8)	Listmonth	Month when the property was listed	Interval
	Listprice (USD)	Price a seller asks	Ratio
	FMR	Fixed mortgage rates	Ratio
	AMR	Adjusted mortgage rates	Ratio
	City	City name	Nominal
	Zip5	Zip code	Nominal
	YearBuilt	Year the property was built	Interval
	DaysOnMarket	Number of days on market	Ratio
Dependent variable	HighOrLow	High: closing price >= listing price Low: closing price < listing price	Nominal

The data integration step involved the combining of the three retrieved datasets into a single cohesive dataset. The Fairfax County public school ratings and weekly mortgage rate tables were integrated into the MLS dataset by matching attributes. First, we integrated the MLS data with the weekly mortgage rates. To achieve this, the "ListDate" attribute in the MLS dataset was matched to the "Weekending" attribute in the mortgage rate table. For example, if the listing date of the property was March 1, 2004, we reviewed the mortgage rate table and found the two entries between which this listing date fell. In this case, the listing date was for the weeks ending February 27, 2004 and March 5, 2004. Therefore, the mortgage rates for the listed property were acquired for the week ending March 5, 2004. Tables 3 and 4 illustrate an example of integrating the MLS dataset with the mortgage rate dataset.

The second dataset to be integrated was the Fairfax County public school ratings. Each school in the MLS dataset was matched with its rating from the public school rating dataset. Three ratings, for elementary, middle, and high school, were added as new attributes; the elementary, middle, and high school name attributes were excluded.

3.3. Data extraction, transformation, and reduction

The first and most important attribute extracted, which subsequently determined the class of each record, was the field indicating if the closing price was higher or lower than the listing price. We called this attribute "HighorLow". It was determined by considering the sign of the difference between the closing price and the listing price. If the difference was positive, then the record became a member of the High class; if the sign was negative, then the record was categorized as Low. Once defined, the closing price attribute was removed from our dataset. The next extracted attribute was the "DaysOnMarket". This was determined by taking the difference between the closing date and the listing date. From the listing date, the listing month was extracted as a new attribute. The listing date and closing date attributes were then removed from the dataset. To reduce the size of the dataset, we excluded all records that had any missing values within the attributes. The final dataset was 5359 records. As illustrated in Fig. 1, the blue color represents a higher closing price (High) and the red color a lower closing price (Low) than the listing price. Of the 5359 records, 3530 records were "High" and 1829 "Low".

3.4. Analysis procedure

In this section, we describe the machine learning procedures we applied to our dataset. In this particular research, we faced one principal challenge: how do we classify a High or Low closing price of a townhouse based on the attributes we have identified. To address this concern, we selected machine learning algorithms with the capability to classify. Four algorithms were chosen: C4.5, RIPPER, Naïve Bayesian, and AdaBoost. For this task, we used the *WEKA* software, a knowledge analysis suite developed by the University of Waikato.

Table 3
An example of the mortgage rate dataset, which contains 30 year FMR, 15 year FMR, and 1 year AMR.

Weekending	30YearFMR[a]	15YearFMR	1YearAMR[b]
02/27/2004	5.58	4.89	3.50
03/05/2004	5.59	4.88	3.47

[a] FMR (fixed mortgage rates).
[b] AMR (adjusted mortgage rates).

Table 4
Integrating the MLS dataset with the mortgage rate table.

ListDate	ListPrice (USD)	30YearFMR[a]	15YearFMR	1YearAMR[b]
02/27/2004	449900	5.58	4.89	3.50
03/01/2004	543515	5.59	4.88	3.47

[a] FMR (fixed mortgage rates).
[b] AMR (adjusted mortgage rates).

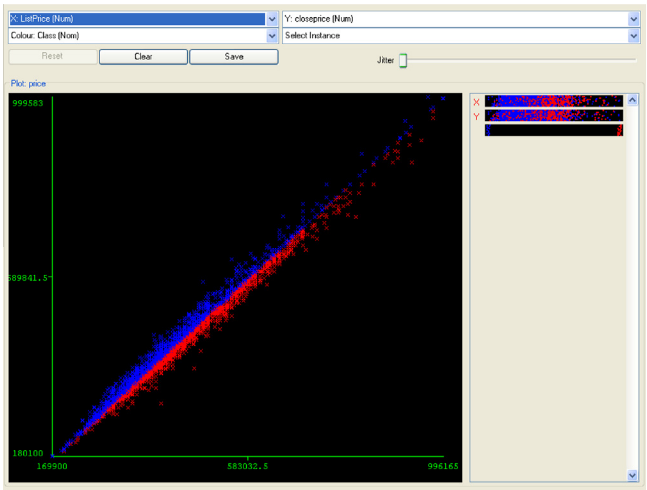

X: ListPrice (Num)
Y: closeprice (Num)
Colour: Class (Nom)
Select Instance

Fig. 1. Visualization of the listing price versus closing price.

A decision tree represents the relationships in a database in a tree structure such that we can trace a path in a tree to classify a new case. J48 is the *WEKA* implementation of the C4.5 decision tree learner (Quinlan, 1993). Of the available decision tree algorithms in *WEKA*, we used the J48 (C4.5). Because the dataset we chose included numerical and nominal attributes, C4.5 was more appropriate, compared to ID3.

Our research question was to determine whether the closing price was higher or lower than the listing price. RIPPER selects the majority class (High class in our case, because there are more records that are of the High class) as its default class and learns the rules for the minority class (Low class). JRip, the propositional rule learner algorithm in the *WEKA* implementation of RIPPER, was used for this research.

Naïve Bayesian is a statistical learning algorithm based on Bayes' rule to compute joint probability. It assumes conditional independence amongst the attributes. This is used as a classification tool by first dividing the dataset into independent classes and calculating the probability distribution for each attribute of each class. For classification, the Naïve Bayesian finds the probability for the unknown in any given class and selects the class with the highest probability.

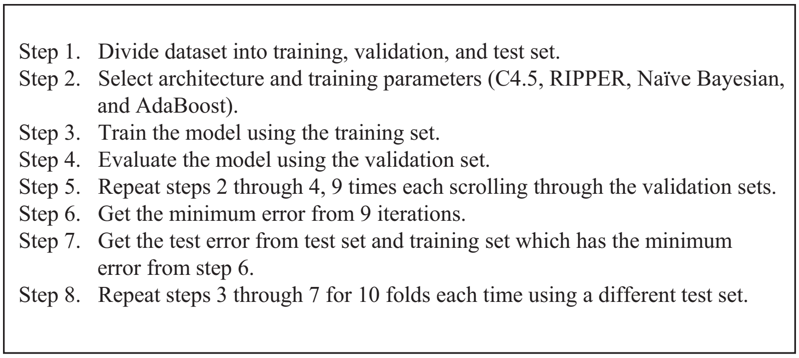
AdaBoost is a method to improve classification using an ensemble of weak classifiers to create a strong classifier. We used AdaBoostM1 as implemented in *WEKA*. AdaBoost functions by repeating rounds of boosting iterations. For each iteration, the dataset is sampled based on the calculated weights and a best weak classifier is found that optimally divides the sampled data into the classes. The chosen weak classifier is then assigned a weight based on how well it divided the data. Classification is performed by obtaining the class for the unknown record for each weak classifier (the weak classifier has a class value of 1 or −1) then multiplying it by the weak classifier's weight and summing. Class is then determined by the sign of the sum.

A systematic method was required to evaluate and compare the classification algorithms to determine the performance for our specific problem. Two methods were explored: three-way split with 10-folds and 10-fold cross-validation. Both of these methods require that the dataset be divided into ten equal parts. Splitting the 5359 records into ten, each portion had 536 records, except for the last, which had 535 records. They were then labeled from 1 to 10. Both evaluation methods used these ten pieces in a slightly different manner. As illustrated in Figs. 2 and 3, the three-way split processed the ten folds where each one included the combination of one test set and one of the remaining nine validation sets. The eight remaining became the training sets. Thus, we had 90 testing combinations for the three-way dataset split. The 10-fold cross-validation, conversely, used only one test set and the remaining nine training sets; therefore, there were a total of ten combinations.

As shown in Fig. 3, the double-loop architecture of the three-way split can be easily seen. The outer loop scrolls through the ten test sets and the inner loop scrolls through the nine validation sets. The remaining is used for training.

4. Experimental results

We applied the two performance tests on the machine learning classifiers. The three-way split was applied to the C4.5, RIPPER, and Naïve Bayesian methods. The 10-fold cross-validation was applied to C4.5, RIPPER, Naïve Bayesian, and AdaBoost. Tables 5–7 display the results for the three-way split. For each fold, the training and validation pairs that returned the minimum error are displayed. The last column is the test error based on the test set of each fold. Fig. 4 presents a plot exhibiting the error rate, with error bars equal to one standard deviation, for the three classifiers analyzed. It can be observed that RIPPER has the smallest error rate and Naïve Bayesian has the highest. Table 8 presents the results for the 10-fold cross-validation. The estimated error rate is shown for each fold and the average and standard deviation are calculated. Fig. 5 presents the error rate, with error bars equal to one standard devi-



Step 1. Divide dataset into training, validation, and test set.
Step 2. Select architecture and training parameters (C4.5, RIPPER, Naïve Bayesian, and AdaBoost).
Step 3. Train the model using the training set.
Step 4. Evaluate the model using the validation set.
Step 5. Repeat steps 2 through 4, 9 times each scrolling through the validation sets.
Step 6. Get the minimum error from 9 iterations.
Step 7. Get the test error from test set and training set which has the minimum error from step 6.
Step 8. Repeat steps 3 through 7 for 10 folds each time using a different test set.

Fig. 2. The three-way data split procedures.

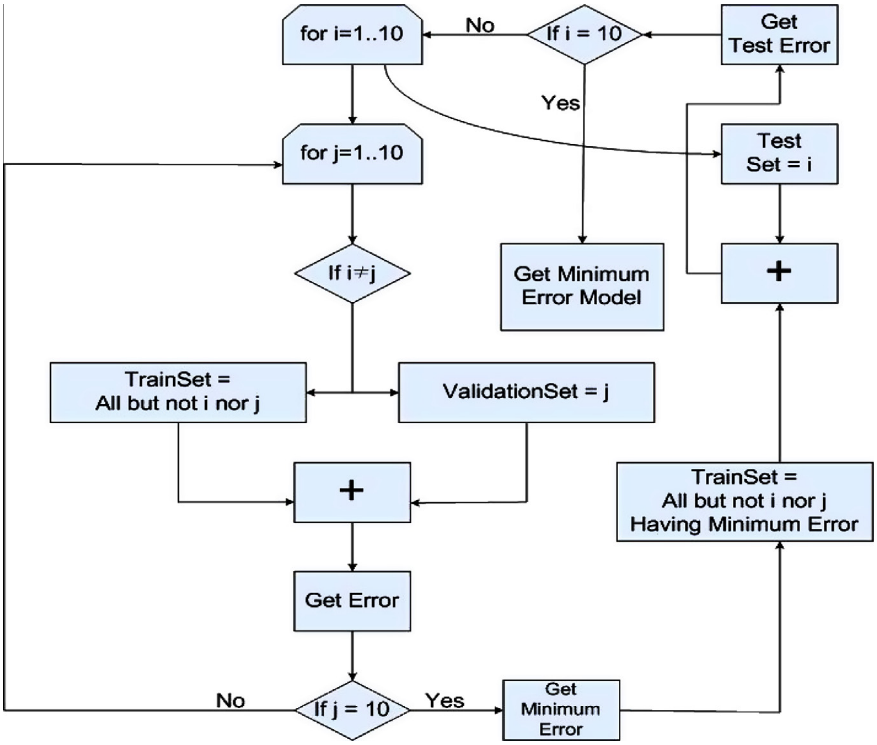


Fig. 3. Flow chart of three-way data split.

Table 5
Three-way dataset test error (C4.5).

Test set (fold)	Validation set	Training set	Minimum fold error	Test error*
1	10	Exclude-1, 10	0.2261	0.2593
2	10	Exclude-2, 10	0.2336	0.2910
3	10	Exclude-3, 10	0.2317	0.2966
4	9	Exclude-4, 9	0.2257	0.2761
5	9	Exclude-5, 9	0.2276	0.2742
6	9	Exclude-6, 9	0.2444	0.2742
7	10	Exclude-7, 10	0.2205	0.3078
8	10	Exclude-8, 10	0.2280	0.2630
9	10	Exclude-9, 10	0.2299	0.2574
10	6	Exclude-10, 6	0.2462	0.2598

* *Note:* Average (test error) = 0.2759, STDEV (test error) = 0.0173.

Table 6
Three-way dataset test error (RIPPER).

Test set (fold)	Validation set	Training set	Minimum fold error	Test error*
1	10	Exclude-1, 10	0.2317	0.2462
2	9	Exclude-2, 9	0.1977	0.2854
3	9	Exclude-3, 9	0.2182	0.2500
4	9	Exclude-4, 9	0.2201	0.2518
5	10	Exclude-5, 10	0.2130	0.2537
6	10	Exclude-6, 10	0.2000	0.2891
7	9	Exclude-7, 9	0.2108	0.2649
8	10	Exclude-8, 10	0.2112	0.2369
9	10	Exclude-9, 10	0.2280	0.2425
10	8	Exclude-10, 8	0.2369	0.2112

* *Note:* Average (test error) = 0.2532, STDEV (test error) = 0.0227.

Table 7
Three-way dataset test error (Naïve Bayesian).

Test set (fold)	Validation set	Training set	Minimum fold error	Test error*
1	4	Exclude-1, 4	0.2817	0.2686
2	1	Exclude-2, 1	0.2686	0.3190
3	1	Exclude-3, 1	0.2686	0.2929
4	1	Exclude-4, 1	0.2686	0.2817
5	1	Exclude-5, 1	0.2705	0.3059
6	1	Exclude-6, 1	0.2667	0.2947
7	1	Exclude-7, 1	0.2723	0.3339
8	1	Exclude-8, 1	0.2705	0.3208
9	1	Exclude-9, 1	0.2742	0.3097
10	1	Exclude-10, 1	0.2779	0.2859

* *Note:* Average (test error) = 0.3013, STDEV (test error) = 0.0201.

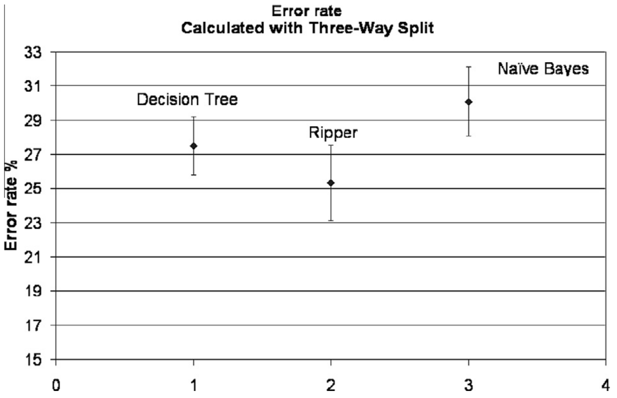


Fig. 4. Error rate, with error bars equal to standard deviation, calculated with the three-way split.

Table 8
Experimental results on 10-fold cross-validation error.

Fold	C4.5	RIPPER	Naïve Bayesian	AdaBoost
1	0.2649	0.2276	0.2705	0.2444
2	0.3134	0.2929	0.3134	0.2966
3	0.2630	0.2593	0.2929	0.2369
4	0.2873	0.2481	0.2779	0.2686
5	0.2630	0.2444	0.3041	0.2537
6	0.2705	0.2537	0.2929	0.2649
7	0.2798	0.2817	0.3283	0.2854
8	0.2481	0.2481	0.3190	0.2667
9	0.2742	0.2238	0.3134	0.2220
10	0.2448	0.2093	0.3009	0.2317
Average error	0.2709	0.2488	0.3013	0.2570
Standard deviation	0.0197	0.0254	0.0181	0.0238

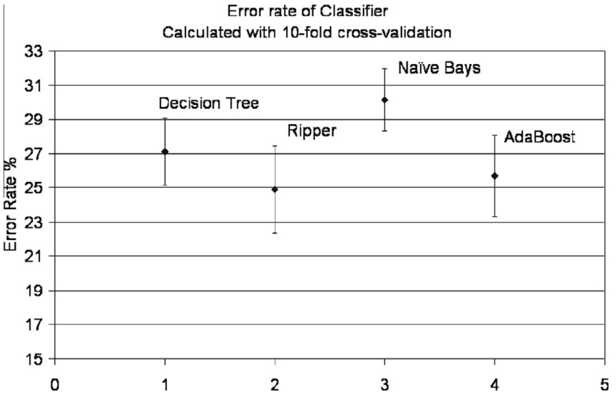


Fig. 5. Error rate, with error bars equal to standard deviation, calculated with 10-fold cross-validation.

ation, for the four classifiers analyzed. It can again be seen that RIPPER has the smallest error rate and Naïve Bayesian has the highest error rate.

5. Conclusions

In this study, several machine learning algorithms are used to develop a prediction model for housing prices. We test for the performance of these techniques by measuring how accurately a technique can predict whether the closing price is greater than or less than the listing price. Four different machine learning algorithms including C4.5, RIPPER, Naïve Bayesian, and AdaBoost are selected, and tested for which algorithm produces the highest rate of the accuracy. We find that the performance of RIPPER is superior to that of the C4.5, Naïve Bayesian, and AdaBoost models. In all the tests, RIPPER outperforms the other housing price prediction models.

Previous studies pertinent to housing price predictions have focused on hedonic-based methods which are conventional statistical approaches having some limitations of assumptions and estimations. More recent research tried to compare conventional ways with machine learning approaches such as neural network and SVM. However, this study compares the performance of various classifiers in machine learning algorithms, and finds the best classifier for a better housing price prediction.

Thus, our study shows that a machine learning algorithm can enhance the predictability of housing prices and significantly contribute to the correct evaluation of real estate price. In practical applications, mortgage lenders and financial institutions can employ a machine learning based housing price prediction model for better real estate property appraisal, risk analysis, and lending decisions. The potential benefits of using this model include reducing the cost of real estate property analysis and enabling faster mortgage loan decisions.

Our study has the following limitations which future research could examine further. This study focuses on a specific region, Fairfax County and on a specific type of residential properties, townhouses. First, location is one of the most important factors in buying and selling real estate because real estate markets have important regional differences (Eichholtz, Veld, & Schweitzer, 2000). Other geographic regions might require different attributes. Second, residential property includes single family houses, townhouses, and condominiums. Results might be different based on the type of residential property. Third, the performance evaluation is based only on classifiers. Performance comparison of other machine learning algorithms should be considered.

In future works, this study can be extended in several ways. First, it could be desirable to investigate other problem domains (real estate market prediction, interest rate forecasting, economic growth rate forecasting, oil price forecasting, and stock price index forecasting) to generalize the results of this study. Second, a future study must establish housing price prediction model that enables forecast of multiclass or continuous dependent variables. Lastly, the housing market can be influenced by macro-economic variables. Future research should consider macro-economic and environmental amenities variables for housing price prediction model inputs. For this purpose, more data sources are needed. For example, property tax and appraised value of a property, and primary residence can be achieved from tax authorities and real estate agency websites.

References

Adair, A., Berry, J., & McGreal, W. (1996). Hedonic modeling, housing submarkets and residential valuation. *Journal of Property Research, 13*(1), 67–83.

Azadeh, A., Ziaei, B., & Moghaddam, M. (2012). A hybrid fuzzy regression-fuzzy cognitive map algorithm for forecasting and optimization of housing market fluctuations. *Expert Systems with Applications, 39*(1), 298–315.

Bin, O. (2004). A prediction comparison of housing sales prices by parametric versus semi-parametric regressions. *Journal of Housing Economics, 13*, 68–84.

Eichholtz, P., Veld, H. O., & Schweitzer, M. (2000). REIT performance. Does managerial specialization pay? In P. T. Harker & S. A. Zenios (Eds.), *Performance of financial institutions: Efficiency, innovation, regulation* (pp. 199–220). New York: Cambridge University Press.

Fan, G., Ong, Z. S. E., & Koh, H. C. (2006). Determinants of house price. A decision tree approach. *Urban Studies, 43*(12), 2301–2315.

Gerek, L. H. (2014). House selling price assessment using two different adaptive neuro-fuzzy techniques. *Automation in Construction, 41*, 33–39.

Gu, J., Zhu, M., & Jiang, L. (2011). Housing price forecasting based on genetic algorithm and support vector machine. *Expert Systems with Application, 38*(4), 3383–3386.

Kauko, T., Hooimeijer, P., & Hakfoort, J. (2002). Capturing housing market segmentation: An alternative approach based on neural network modeling. *Housing Studies, 17*(6), 875–894.

Kestens, Y., Theriault, M., & Rosier, F. D. (2006). Heterogeneity in hedonic modelling of house prices: Looking at buyers' household profiles. *Journal of Geographical Systems, 8*(1), 61–96.

Kuşan, H., Aytekin, O., & Özdemir, İ. (2010). The use of fuzzy logic in predicting house selling price. *Expert Systems with Applications, 37*(3), 1808–1813.

Liu, J., Zhang, G. X. L., & Wu, W. P. (2006). Application of fuzzy neural network for real estate prediction. *LNCS, 3973*, 1187–1191.

Meese, R., & Wallace, N. (2003). House price dynamics and market fundamentals: The Parisian housing market. *Urban Studies, 40*(5–6), 1027–1045.

Quinlan, J. R. (1993). C4.5: Programs for machine learning, San Mateo, CA: Morgan Kaufmann.

Schulz, R., & Werwatz, A. (2004). A state space model for Berlin House prices: Estimation and economic interpretation. *Journal of Real Estate Finance and Economics, 28*(1), 37–57.

Selim, H. (2009). Determinants of house prices in Turkey: Hedonic regression versus artificial neural network. *Expert Systems with Applications, 36*(2), 2843–2852.

Stevenson, S. (2004). New empirical evidence on heteroscedasticity in hedonic housing models. *Journal of Housing Economics, 13*, 136–153.

Wang, X., Wen, J., Zhang, Y., & Wang, Y. (2014). Real estate price forecasting based on SVM optimized by PSO. *Optik-International Journal for Light and Electron Optics, 125*(3), 1439–1443.