House Price Prediction

Data Mining technique that predeict the price value of the house.

1st Hebah A. Alshamlam
dept. of Computer Science
(43720117)

2nd Maha S. Alqadeeb
dept.of Information System
(437200224)

3rd Sarah S. Alwabel
dept.of Information System
(437202023)

4th Munerah A. Aljbreen
dept.of Information System
(436202227)

***Abstract*— An effective estimation of a house price between sales is needed. A house sale or real estate agent can be aided in making more educated choices depending on the house prices. This paper attempts to demonstrate how house pricing can be estimated by the data mining techniques. We also conclude, on the basis of the studies performed in this paper, that SVM model is the better for this problem.**

Keywords— Prediction, Data Mining, Linear Regression, Support Vector Regression.

I. Introduction

Through decades the real estate industry has been growing and increasing in size dramatically, and a significant change through years been noticed in prices. We know for a fact that many factors have caused this effect on the price and factors such as interest rate, GDP, location...etc. Since there is a growth in the industry, this means people are still buying properties, and there is a need in this field. Therefore, a small observation has been made, and we found there is a minor implementation of technology in the industry. From here, the idea of House Price Prediction showed up, using big data and ML, and there are several benefits of price prediction if we talked about governmental perspective it is a crucial factor in the urban development, will help them to do rational planning for it, and if we took a closer look at the homeowners, based on General Authority Statistic there[1] are 5.46million (2016) homeowners, and if we assumed four people per household, this means around 70% of the population are homeowners. Therefore, predicting prices will help them build long-term strategies and invest in the most utilized and profitable industry and this will lead to economic stability. In conclusion, House Price Prediction will provide a verity of benefits for both government and citizens and will illustrate the industry in general.

II. Literature Review

This section will provide some literature review and an overview of how data mining techniques were applied to predict house prices. The authors [2] aim to build a system that predicts house prices and future prices for real state customers using the classification technique and Naïve Bayes will be used to divide data into independent classes and calculate the probability distribution for all the attributes. They address this problem with the approach of analyzing current and previous market prices and features to estimate the prices. Also, the author Nguyen[3] explores the question of how house prices in five different counties are affected by housing characteristics (both internally, such as several bathrooms, bedrooms, etc. furthermore externally, some of which are public schools' scores or the walkability score of the neighborhood). They used a dataset of 1,457 houses from 5 different counties scraped from Zillow, Trulia, and Redfin. The models experimented are Linear Regression and various machine learning algorithms, such as Random Forest (RF) and Support Vector Regression (SVR). The results show that SVR gives a better price prediction score than the Zillow's baseline on the same dataset This paper identifies the four most important attributes in housing price prediction across the counties as assessment, comparable houses' sold price, listed price and number of bathrooms. And the result of Support Vector Regression Gives a Ratio of 1:1. On the other hand, Stephen[4] examined the efficiency of data mining techniques toward the human-influenced dependent variable. The paper used regression techniques with six algorithms such as Linear regression, K- nearest neighbor, and the support vector machine. The paper uses three different metrics to evaluate the performance of each algorithm the Root Mean Square Error (RMSE) is the standard deviation of the prediction errors, Mean Absolute Error (MAE) measures the average of all absolute errors, and R-Squared (R2) which represents a variance ratio in a regression model. The linear regression and the partial least squares have an RMSE equal to 0.19 and MAE is 0.14 and high in R2 is 0.87. These algorithms have achieved the lowest RMSE and MAE values and highest R-squared values of the five algorithms. The stacked model created similar results compared to multiple linear regressions and partial least squares. In the end, there is a strong correlation between the model's predictions and the actual observed house prices. Also The authors [5] aims to build Regression model and Classification model that can accurately estimate house prices given its features. Using Lasso, Ridge, Support Vector Machine (SVM) regression, and Random Forest regression as regression algorithms for predicting continues house prices. And Naive Bayes, logistic regression, Support Vector Machine (SVM) classification, and Random Forest classification as classification methods to predict individual price ranges. Moreover, used Principal Component Analysis (PCA) to improve the accuracy of the prediction. The dataset has the data of houses in Ames, Iowa. And have 79 different features to predict based on, and 1460 records. Results of the classification problem showed that the best performing model is the SVM classification with linear kernel, an accuracy of 0.6740. However, with PCA performed, the accuracy increased to 0.6913. Whereas results of the regression problem showed that the best performing model is SVM regression with root mean square error of 0.5271. Finally, the authors in [6] aim to predict the house pricing by analyzing current house prices thus forecasting future prices. They think

that the regular method of approaches a real estate agent to suggest suitable estates for investment is outdated and has high risk as the agent might predict the wrong estates and thence leading to loss of the customer's investments, to overcome this fault, they implement the linear regression algorithm so they can accurately forecast house prices. The dataset they used contains real estate prices in Navi Mumbai, India from January 2009 to December 2015 and where each year is divided into 4 quarters (q1: January-March, q2: April-June, q3: July- September, q4: October-December), It also is divided into 3 categories namely; Upper, it represents the houses by renowned builder associations and full of amenities followed by subsequent fewer categories as Average and Lower. The result they came up with is that the linear regression algorithm helps to satisfy customers by increasing the accuracy and efficiency of estate choice and decrease the risk of investing in an estate.

III. DATASET

a) Dataset Description

This project will be based on the House Price Prediction dataset found in *Kaggle* [7]. It is based on house price sales during the second quarter of 2014 in various cities in Washington, USA. Overall, the dataset contains 4,600 instances and 18 attributes with the price as a class label. A description of the data set is provided in Table 1.

TABLE 1 DESCRIPTION OF DATASET

Attribute	***Description***	***Type***
Date	This is the date the house is added into the dataset	Date
Price -Class Label	Price of each house in US dollars.	Numeric
Bedrooms	The number of bedrooms available in each house	Numeric
Bathrooms	The number of bathrooms available in each house	Numeric
Floors	The number of floors available in each house	Numeric
Waterfront	An indicator if the house is located on or beside a lake or beach. (0,1)	Numeric
Yr_bulit	The year of a house in which it is constructed.	Numeric
Yr_renovated	Year in which the house is renovated or remodeled.	Numeric
Sqft_lot	The total area of the size of lot in square feet.	Numeric
Sqft_living	Area size of the living room in square feet.	Numeric
Sqft_above	The surface area of house in square feet above ground level.	Numeric
Sqft_basement	The surface area of house in square feet below ground level or basement.	Numeric
View	Rating of view of city or lake or beach from the house and is rated from 0 to 5.	Numeric
Condition	Overall condition of house rated in the range 1 to 5.	Numeric
Street	Name of the street in which the house is located.	Nominal
City	Name of the city in which the house is located.	Nominal
StateZip	A 5-digit zip code in which the house is located.	Nominal
Country	Name of the country in which the house is located.	Nominal

b) Histogram

Histogram reviews the graphical distribution of each attribute. The first histogram shows the distribution of the numeric type of the attribute *price*. This attribute has no missing values with the range between 80,000 to 2,888,00 as shown in figure 1.

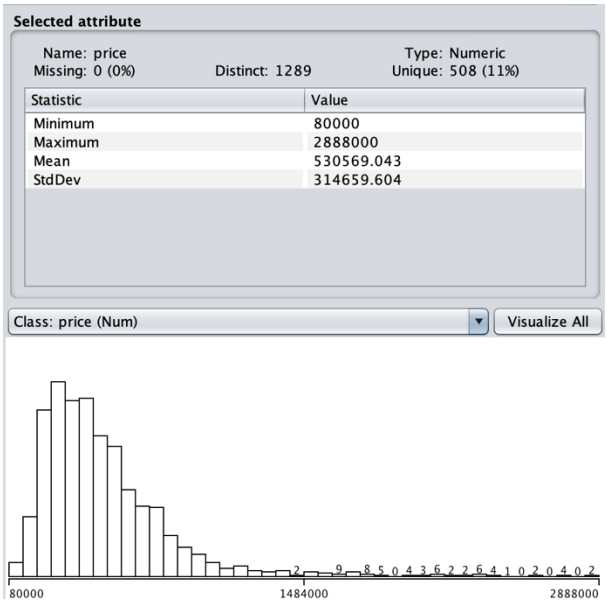


FIGURE 1 PRICE DESCRIPTION AND HISTOGRAM

The next histogram shows the distribution of the nominal type of the attribute *yr_build*. This attribute has no missing values and a range of numeric attributes in the dataset into nominal attributes as shown in figure 2.

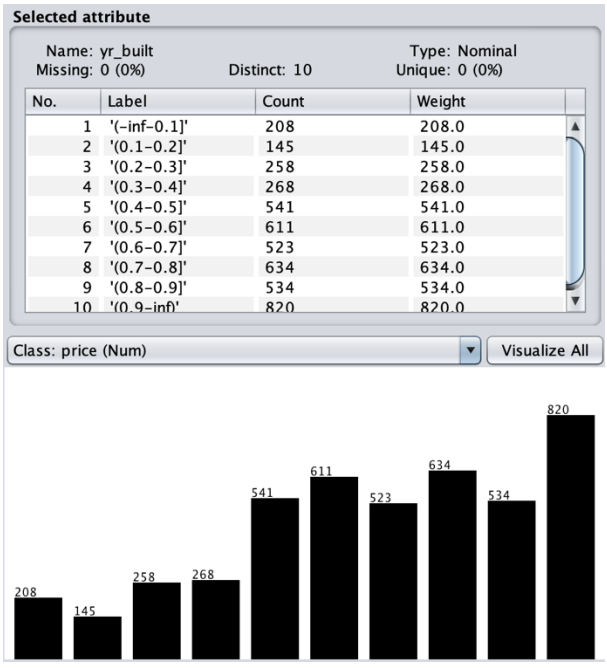


FIGURE 2 YR_BUILD DESCRIPTION AND HISTOGRAM

IV. METHODOLOGY

An important application of Artificial intelligence is Machine Learning (ML) which is uses algorithms and statistical models with the ability to learn and improve from training and experience. Machine Learning is a powerful approach. In this section we will discuss our pipeline including preprocessing and ML models.

A. Correlation

To consider the relationships between the attributes by using the attribute evaluator *CorrelationAttributeEval* and *Ranker* as a search method. All the attributes have a value higher than zero which admits as a positive correlation and there is not a negative correlation in the dataset. The most correlated attribute with the *price* is the *sqft_living* with a 0.6904. Figure 3 shows the correlation between the attributes.

```
=== Attribute Selection on all input data ===

Search Method:
        Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 15 price):
        Correlation Ranking Filter
Ranked attributes:
 0.6904   3 sqft_living
 0.5896   9 sqft_above
 0.5095   2 bathrooms
 0.331    7 view
 0.3175   1 bedrooms
 0.2673   5 floors
 0.1881  10 sqft_basement_binarized_binarized
 0.1487   6 waterfront_binarized_binarized
 0.1007   4 sqft_lot
 0.0916  13 city
 0.0634  14 statezip
 0.0487   8 condition
 0.0292  12 yr_renovated_binarized_binarized
 0.0152  11 yr_built

Selected attributes: 3,9,2,7,1,5,10,6,4,13,14,8,12,11 : 14
```

FIGURE 3 CORRELATION RANKING

B. Preprocessing

Before begin modeling it is necessary to review the data to discover the importance of each attribute and the correlations between attributes. In the beginning, we decided to remove three attributes the date, street, and country.

Moreover, the predictor variable *yr_renovated* contains mostly zero fields which specified the house has not renovated. To handle the previous issue the attribute type had been changed to binary. Likewise, the *sqft_basement* and the *waterfront* have been changed to binary. This process is performed by the filter weka.filters.unsupervised.attribute .NumericToBinary. Besides, there were some instances with *price* equal to zero which is not acceptable so they were removed.

The next step was to normalize all the numeric variables to the range of 0 to 1 this will change the mean and the standard deviation as well to be applied to the new range as shown in figure 4. This process is performed by the filter weka. filters.unsupervised.attribute.Normalize.

Name: sqft_living Type: Numeric
Missing: 0 (0%) Distinct: 554 Unique: 175 (4%)

Statistic	Value
Minimum	370
Maximum	13540
Mean	2124.429
StdDev	931.882

Name: sqft_living Type: Numeric
Missing: 0 (0%) Distinct: 554 Unique: 175 (4%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.133
StdDev	0.071

FIGURE 4 APPLYING NORMALIZATION

Next, the attributes *yr_built* have been discretized by simple binning. this filter decreases the distinct values from 155 to 10 as shown in figure 5. This process is performed by the filter weka.filters.unsupervised.attribute.discretize.

Name: yr_built Type: Numeric
Missing: 0 (0%) Distinct: 115 Unique: 0 (0%)

Statistic	Value
Minimum	0
Maximum	1
Mean	0.621
StdDev	0.261

Name: yr_built Type: Nominal
Missing: 0 (0%) Distinct: 10 Unique: 0 (0%)

No.	Label	Count	Weight
1	'(-inf-0.1]'	205	205.0
2	'(0.1-0.2]'	172	172.0
3	'(0.2-0.3]'	255	255.0
4	'(0.3-0.4]'	270	270.0
5	'(0.4-0.5]'	539	539.0
6	'(0.5-0.6]'	601	601.0
7	'(0.6-0.7]'	507	507.0
8	'(0.7-0.8]'	623	623.0
9	'(0.8-0.9]'	529	529.0

FIGURE 5 DISCRETIZING FILTER

The last preprocessing filter was resampling to produces a random subsample of a dataset as shown in figure 6. This process is performed by the filter weka.filters.supervised .instance.Resample.

Name: city Type: Nominal
Missing: 0 (0%) Distinct: 44 Unique: 3 (0%)

No.	Label	Count	Weight
1	Tukwila	29	29.0
2	Carnation	22	22.0
3	Covington	42	42.0
4	Yarrow Point	4	4.0
5	Auburn	175	175.0
6	Seattle	1559	1559.0
7	Kirkland	187	187.0
8	Kent	183	183.0
9	Renton	291	291.0

Name: city Type: Nominal
Missing: 0 (0%) Distinct: 43 Unique: 3 (0%)

No.	Label	Count	Weight
1	Tukwila	25	25.0
2	Carnation	17	17.0
3	Covington	55	55.0
4	Yarrow Point	2	2.0
5	Auburn	186	186.0
6	Seattle	1536	1536.0
7	Kirkland	179	179.0
8	Kent	182	182.0
9	Renton	299	299.0

FIGURE 6 RESAMPLING FILTER

C. Regression Technique

Regression is a data mining technique used to predict a range of numeric values, given a particular dataset. Regression can be used to model the relationship between one or more independent variables and dependent variables. In data mining, independent variables are attributes already known and response variables are what we want to predict. Some of the regression algorithms are Linear regression, Multiple Regression Algorithm, Logistic regression and Support Vector Machines. Since regression technique is very useful in

predicting numeric values it has been used in this paper to predict house pricing.

D. Models

1) Linear Regression

One of the most commonly used algorithms in regression technique is linear regression. It is a linear method that model the relationship between variables. Also, used to predict value of the dependent and continues variable Y based on the values of an independent variable X, and it could be either continues or discrete values. Using the following equation to represent multiple linear regression model where Y is the response or outcome variable and x is predictor variables:

$$\boldsymbol{Y = b_0 + b_1 x_1 + b_n x_n}$$

One advantage of this model is that it is considered the least complex linear model in finding relation between independent and dependent variables. Additionally, it selects the optimal model through the least square criterion which reduces the squared error between expected and actual values.

2) Support Vector Regression

A commonly used "off-the-shelf" supervised learning approaches is support vector machine or SVM, the reasons of its popularity are: It provides maximum margin separator, which has a decision boundary that has the largest distance between example points; this helps to generalize well. It also uses kernel trick, which embeds data into higher dimensional space to be easily separable if the data is overlapping. SVMs combine advantages of nonparametric and parametric models; complex functions can be represented easily, without resulting in overfitting [8]. Figure 8 (a) shows a two dimensional training set with negative examples (white) and positive example (black). Figure 8 (b) shows the same training data after mapping it into a higher dimension (three dimensions), now we can apply linear decision boundary.

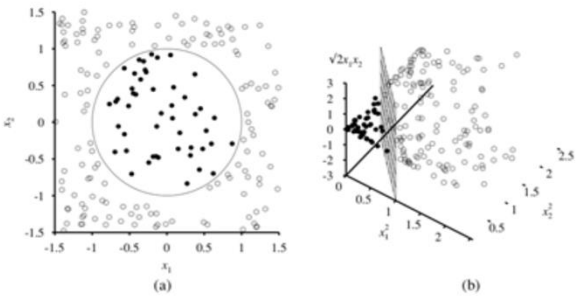
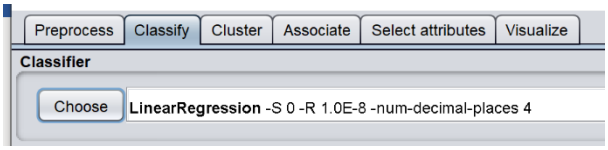


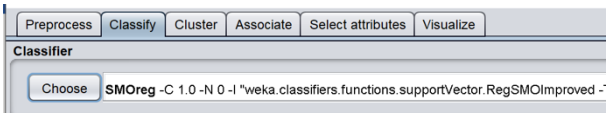
Figure 7 Decision boundary made by SVM.

V. Implementation

Chose *LinearRegression* from Classify tap to run the model



For Support Vector Regression Chose *SMOreg*



some parameters were modified. *C = 3* and the chosen kernel is *Puk.*

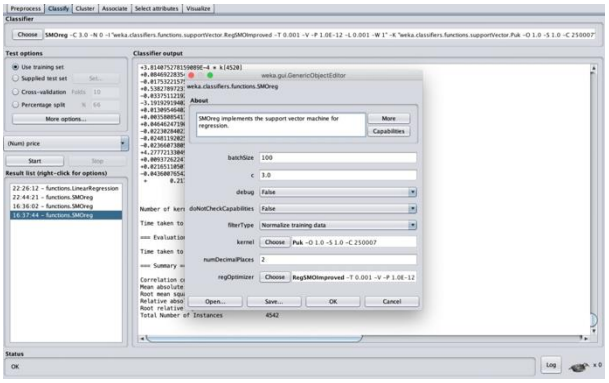


FIGURE 8 DEFINING PARAMETERS FOR SVR

VI. Evaluation

In this section, we will show the experimental tables we go through to tune the parameters. We did parameter-tuning for Support Vector Regression. Also, the pipeline of the preprocessing steps.

A. Preprocessing pipeline results

To achieve the ideal dataset preprocessing some experiments has been done as shown in table 1. The table shows how accuracy is improved when adding a new preprocessing step. At the end of these experiments, test#5 has the highest accuracy in linear regression and has been adopted. Also, because there is no parameter we can change on Linear Regression so we test the dataset on this model.

TABLE 2 PREPROCESSING PIPELINE

Test No.	***Preprocessing Steps***	***Accuracy in LR***
Test#1	Data cleaning (deleting attributes + convert numeric to binary)	55.18%
Test#2	Data cleaning (deleting attributes + convert numeric to binary) + normalization	55.24%
Test#3	Data cleaning (deleting attributes + convert numeric to binary + Deleting outliers from price) + normalization	88.63%
Test#4	Data cleaning (deleting attributes + convert numeric to binary + Deleting outliers from price) + normalization + Discretization	88.95%
Test#5	Data cleaning (deleting attributes + convert numeric to binary + Deleting outliers from price) + normalization + Discretization + Resample	89.03%

B. Experimental table of Support Vector Regression

Support Vector Regression has two main parameter Kernel and C. And the table below display the tuning has been made

TABLE 3 SVR EXPERIMENTAL TABLE

Test No.	*c*	*kernel*	*Accuracy*
Test#1	1	Puk	99.86%
Test#2	2	PolyKernel	88.57%
Test#3	2	RBFKernel	89.6%
Test#4	2	Puk	99.94%
Test#5	3	PolyKernel	88.58%
Test#6	3	RBFKernel	90.14%
Test#7	3	Puk	99.96%
Test#8	2	NormalizedPolyKernel	94.74%
Test#9	3	PolyKernel	88.58%
Test#10	3	RBFKernel	90.14%

VII. RESULT

```
=== Summary ===

Correlation coefficient                  0.8903
Mean absolute error                  89115.6921
Root mean squared error             143256.546
Relative absolute error                 41.0871 %
Root relative squared error             45.5325 %
Total Number of Instances             4542
```

FIGURE 9 RESULT OF LINEAR REGRESSION

```
=== Summary ===

Correlation coefficient                  0.9996
Mean absolute error                   3720.2059
Root mean squared error               9466.7288
Relative absolute error                  1.7152 %
Root relative squared error              3.0089 %
Total Number of Instances             4542
```

FIGURE 10 RESULT OF SUPPORT VECTOR REGRESSION

Below are the result of the RMSE, MAE, and Accuracy of the two models we explored. We see that the SVR gives the lowest RMSE and MAPE, Also, a higher accuracy than LR. However, SVM attempts to find the "right" margin dividing the groups, and this decreases the probability of error on the results, while linear regression does not, rather it will have different judgment boundaries with different weights close to the optimum level. Moreover, There's less chance of overfitting in SVM, though linear regression is vulnerable to overfitting.

VIII. CONCLUSION

In this paper, regression techniques were used to predict house prices based on several attributes on the chosen dataset. linear regression and support vector regression were used as regression algorithms. Moreover, the experiments and results of both algorithms was discussed in detail. Finally, we conclude based on the previously discussed results that SVR has a better result in predicting house prices than linear regression.

As future work, it will be interesting to explore deep learning techniques such as Long Short-Term Memory (LSTM) it is for long sequences problem. So, it could give better results. On the other hand, it will be interesting to use a large Saudi dataset with recent data.

ACKNOWLEDGMENT

In the name of Allah, the Most Gracious and the Most Merciful, our deepest gratitude to the people who contributed in this field and made our work essay to complete. For those involved in data mining who have made daily life more productive and less time consuming.

REFERENCES

[1] "Demography survey", Stats.gov.sa, 2020. [Online]. Available: https://www.stats.gov.sa/sites/default/files/en-demographic-research-2016_2.pdf .

[2] N. Bhagat, A. Mohokar, and S. Mane, "House Price Forecasting using Data Mining," Semantic Scholar, 01-Jan-1970.

[3] An, Nguyen, Chris Fernandes, N. Webb and Harlan Holt. "Housing Price Prediction." (2018).

[4] S. O'Farrell, "Comparison of Data Mining Models to Predict House Price", Academia.edu, 2018.

[5] H. Yu and J. Wu, "Real Estate Price Prediction with Regression and Classification", *Cs229.stanford.edu*, 2020. [Online]. Available: http://cs229.stanford.edu/proj2016/report/WuYu_HousingPrice_report.pdf.

[6] "House Price Forecasting using Data Mining Techniques", International Journal of Advanced Research in Computer and Communication Engineering, vol. 6, no. 12, pp. 1-10, 2020.

[7] Kaggle.com. 2020. House Price Prediction. [online] Available at: https://www.kaggle.com/shree1992/housedata .

[8] Stuart J. Russell , Peter Norvig, "Artificial intelligence: a modern approach", Prentice-Hall, Inc., Upper Saddle River, NJ, 1995