House Price Prediction

Arjun S M	Kartik Shah	Madhu Vamsi	Srilatha Bekkem
200199318	200176617	200203628	200204586
asivaga@ncsu.edu	kcshah@ncsu.edu	mmachav@ncsu.edu	sbekkem@ncsu.edu

1 Background

1.1 Problem

House price prediction is significant to developers, prospective house owners, investors, appraisers, tax assessors, mortgage lenders and insurers. Therefore, an accurate prediction of real estate trends and prices assists local governments and companies make informed decisions. Traditionally, the price of the house is predicted using cost and sale price comparison, but this model lacks an accepted standard. A recent report shows that house sellers and buyers are more inclined towards researching online in order to predict the price of the house before contacting real estate agents. The objective of the project is to build up a model that has the capability of automated machine learning that can do analysis and estimate the price of the house that is up for sale. We will study and comprehend the impact of important factors including house size, house age, number of bedrooms, number of bathrooms and geographical location that influence the pricing of the houses.

There are several factors that affect prediction of prices of houses. Over the last two decades there has been a proliferation of empirical studies analyzing residential property values, with Ball (1973) being last major study. Each succeeding research has generally improved the predictive power of the models by emphasizing attributes of property value such as housing site, housing quality, geographical location and the environment. More recent studies have focused on location externalities, transaction costs and factors affecting the future expected cost in home ownership.

1.2 Literature Survey

In one of the research, the factors are divided into three main groups, there are physical condition, concept and location. Physical conditions are properties possessed by a house that can be observed by human senses, including the size of the house, the number of bedrooms, the number of bathrooms, the availability of waterfront and the age of the house, while the concept is an idea offered by developers who can attract potential buyers, for example, the concept of a minimalist home, healthy and green environment, and elite environment. Location is an important factor in shaping the price of a house. This is because the location determines the prevailing land price. In addition, the location also determines the ease of access to public facilities, such as schools, campus, hospitals and health centers, as well as family recreation facilities such as waterfronts and malls.

2 Proposed Method

2.1 Intuition

The project will incorporate dealing with missing information, information cleaning to demonstrate more powerful analysis. In addition, to assess the data set, we will be using different regression models. The number of attributes that influence the price of the house is large, and this makes it difficult for an individual to decide how much a house is worth without a model. Our model will help fill the gap and improve the efficiency of house price prediction. Our research uses machine learning algorithms to develop a housing price prediction model. This undertaking will be helpful to decision makers such as land specialists, clients, real estate agents, in light of the fact that the multivariate

analysis will locate the best combination to foresee the estimation of the house which is available on sale dependent on the qualities.

2.2 Description of Algorithms

2.2.1 Linear regression

Linear regression is a linear approach to model the relationship between a dependent variable and one or more explanatory independent variables. Linear regression tries to establish the relationship between dependent and independent variables as a model by fitting a linear equation to the observed data. The independent variable is also called as the target variable or the explanatory variable. The explanatory variable in our case is the price of the house. Before implementing linear regression and trying to fit a linear regression to the observed data, we should ensure that there exists some type of association between the dependant and explanatory variables. Also, we have used a scatter plot to show the strength of the relationship between the dependent and explanatory variables.

2.2.2 Random Forest for regression

Since here the given data set has a continuous output, so we utilize regression trees. The algorithm calculation fills in as a vast collection of non correlated decision trees. It makes a lot of decision trees and utilize them to settle a decision. The regression trees are selected to minimize either 1) Variance (The split with lower variance difference is chosen as the criteria to part the values) or 2) Mean absolute error within all subsets. This is a method dependent on bagging and the trees in random forests keep running in parallel with no cooperation with one another.

2.2.3 K-Nearest Neighbour Classifier with Principal Component Analysis

KNN is used for both regression and classification problems. It is easy to interpret output and has low calculation time. This model predicts the class value of data point by considering the average of the K nearest neighbors i.e. K nearest data point's target value to this data point. Euclidean, Manhattan or Hamming are few methods of calculating distance between points.

Principal Component Analysis or PCA is a technique to reduce the dimensions of a data set. There are some attributes of data which contribute a lot to the variance compared to other. PCA linearly transforms the coordinate system to a direction which captures the max variance and thus we can eliminate some attributes which don't have very less variance.

2.2.4 Deep Learning using Artificial Neural Networks

One of the data mining techniques that is used for classification and clustering is Artificial Neural Networks. ANN memorizes every single call. Hence, it is a machine learning technique with enormous memory. It has many different coefficients, which can be optimized. ANN model consists of three layers namely input layer, hidden layer and output layer. Input layer nodes are connected to hidden layer nodes, hidden layer nodes are connected to output layer nodes. One of the mechanisms to correct the weights is through back propagation with gradient descent.

3 Plan and Experiments

3.1 Data Set

3.1.1 Description of the Data Set

The data set used for the project is taken from Kaggle, and it includes details about House Sales in King County, United States of America.

`https://www.kaggle.com/harlfoxem/housesalesprediction`

The data set consists of 19 house features along with price and id columns. These features can be divided into six categories. They are,

Room related: number of bedrooms and number of bathrooms

Size of the house: sqft_living, sqft_above, sqft_lot, sqft_basement, sqft_living15, sqft_lot15 and floors

Date: date, yr_built and yr_renovated

Geographical Location: latitude, longitude and zip_code

Rating: condition and grade

Attractions: view and waterfront

Also, the data set consists of 21,613 records with no missing or duplicate data.

3.1.2 Pre-processing

The data set used for our project contains 'zipcode' and 'id' which we felt is not correlated to the target attribute 'price'. So, these attributes were dropped from both training and testing data while constructing the models. A new attribute "house_age" using "year_built" attribute of the data set is considered while developing the models. Also, "latitude" and "longitude" attributes consist of very large numbers while others are small numbers such as "grade", "number of bedrooms" etc, so we performed standardization on all attributes. The data set was checked for missing data and no missing or duplicate data were found. Further, for all the models implemented, the data set was divided into 80% training data and 20% testing data.

3.2 Details of the Experiments

3.2.1 Linear regression

Linear regression is one of the most commonly used types of predictive analysis. Linear regression focuses on how good the predictor variables detect the outcome of the explanatory variable, and it picks out the subset of predictor variables that has the major impact on the outcome of the explanatory variable. There are various types of linear regression. They are, 1) Simple linear regression 2) Multiple linear regression 3) Logistic regression 4) Ordinal regression 5) Multinominal regression 6) Discriminant analysis

We have used simple linear regression for our analysis.

3.2.2 Random Forest for regression

The algorithms works in such a way that we get a sample set and from this sample set we create a ton of subsets with random values. From these arbitrary subsets we make distinctive choice trees. After getting all the decision trees, whenever given another component, we have to get the predicted value by asking all the decision trees. And afterward whatever vote is most elevated we view it as the estimation of it. Also, the general strategy is to minimize the error in each leaf. We will also find what features positively affect the house costs.

3.2.3 K-Nearest Neighbour Classifier with Principal Component Analysis

Reducing the number of components can be very helpful in KNN distance calculations as higher the number of attributes the more the computation and time complexity required for the model to process the test data. We will choose the optimal K factor by finding the minima in the elbow plot generated by the results of KNN algorithm by varying K value.

Working of KNN Model:

1 Initialize the value of K

2 To find predicted class, iterate through training data points

2.1 Calculate distance between test data and each row of training data.

2.2 Sort the data in ascending order based on distance values

2.3 Consider top K rows and calculate the average of these rows

2.4 Return the predicted class

3.2.4 Deep Learning using Artificial Neural Networks

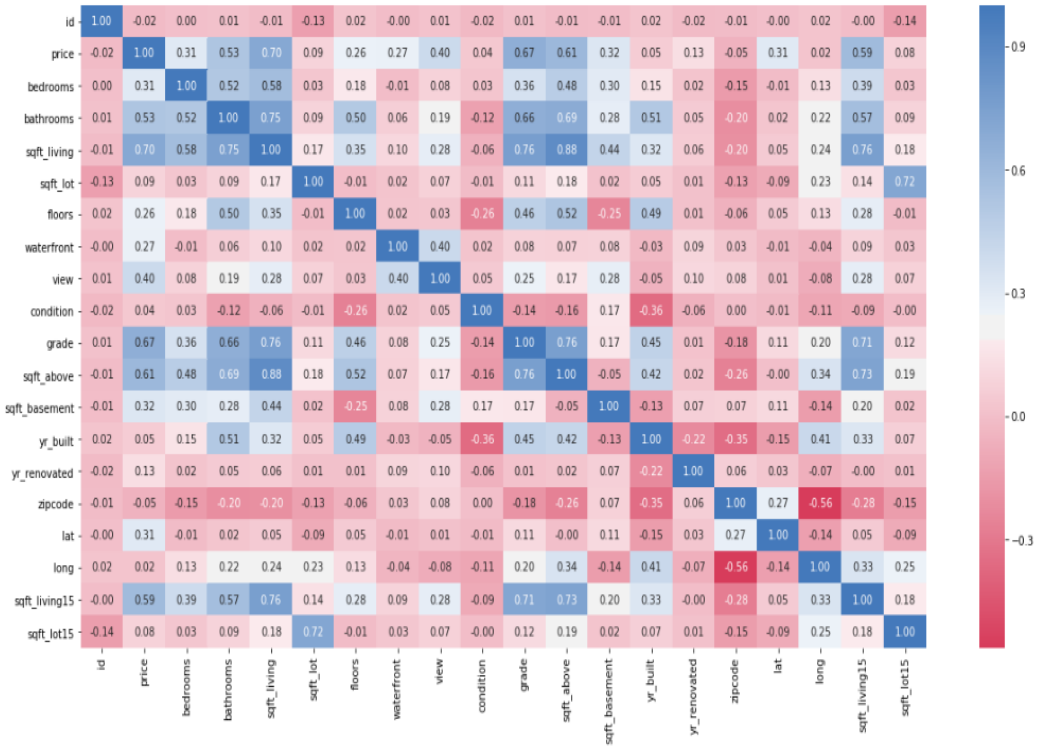
Working of ANN model:

1. Standardize the data.
2. Divide data into training and testing dataset.
3. Assign Random weights to all links.
4. Decide on epoch, hidden layers and neurons
5. Find the activation rate of hidden nodes using inputs then find activation rate of output nodes using hidden nodes.
6. Find the error rate at output node.
7. Re-calibrate between hidden nodes and output nodes
8. Cascade down to hidden nodes and input nodes.
9. Repeat the process till convergence point is met.
10. Use the final weights score the activation nodes of output nodes.

4 Results

4.1 Linear Regression

Results: We obtained three results from our Linear Regression model. One of the important reasons for using Linear Regression model is to find which predictor attribute has the major impact on the target variable. So, we implemented a heat map for the given data to identify the attributes that are highly correlated to the price attribute.



Critical Evaluation: From the heat map, we observed that sqft_living and grade have the maximum correlation with the price attribute. So, we decided to model the relationship between price and these attributes. Unfortunately, the models with sqft_living and grade gave accuracy of 49% and

48% respectively. Ultimately, we decided to model using all the predictor attributes available and the accuracy of the final model was found to be 70%.

4.2 Random Forests

Results: Random forest is used for regression by constructing a multitude of decision trees at training time and outputting the class i.e. mean prediction (regression) of the individual trees. We tried to implement random forest with different number of individual trees. The results we arrived at are as follows,

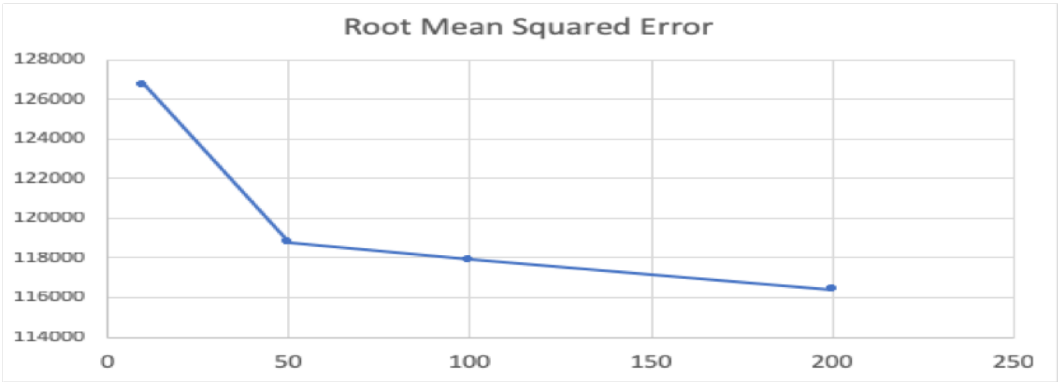
Ist Model: Number of estimators = 10 => The root mean squared error = 68822.52

2nd Model: Number of estimators = 50 => The root mean squared error = 64391.26

3rd Model: Number of estimators = 100 => The root mean squared error = 63848.35

4th Model: Number of estimators = 200 => The root mean squared error = 63701.22

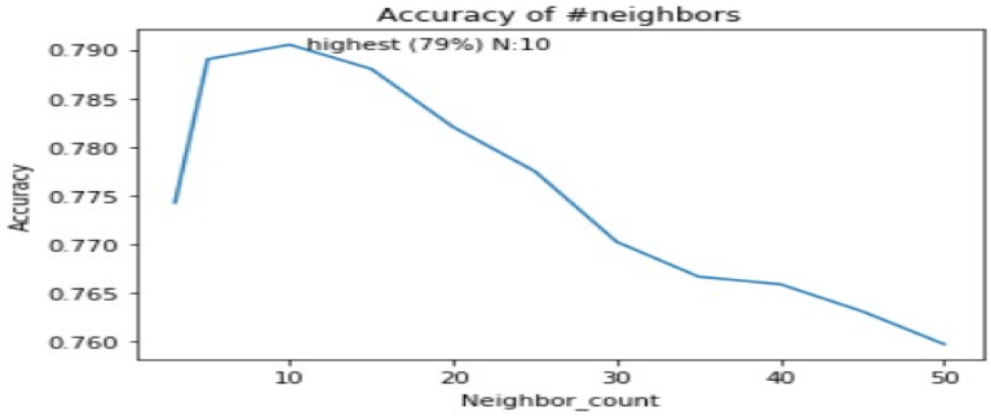
We plotted the increase in the accuracy of the model with the increase in the number of estimators in the form of a graph.



The accuracy of Random Forests model was found to be 88%

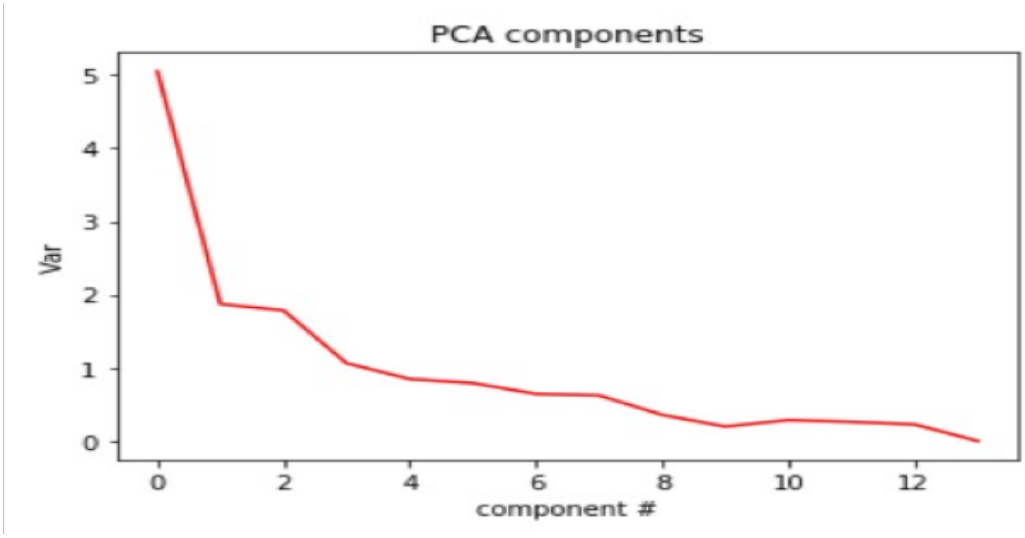
4.3 K-Nearest Neighbour Classification

Results: In K-Nearest Neighbour Classification, the output is the average of the values of its k nearest neighbours. We performed the experiment for various values of k [3,5,10,12,...,50] and plotted the graph.

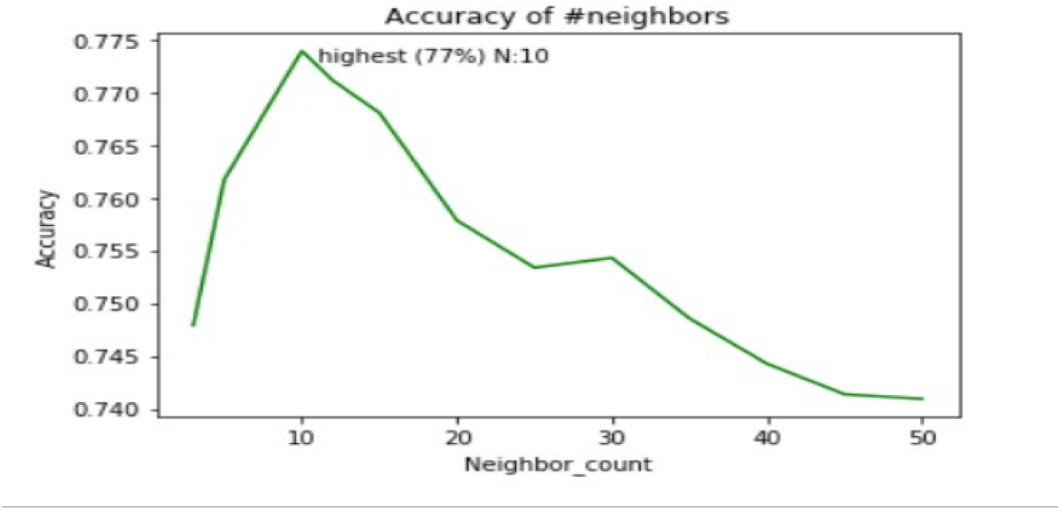


The maximum accuracy achieved is 79%

Principal Component Analysis is a dimensionality reduction procedure that uses an orthogonal transformation to convert a set of observations into a set of values of linearly uncorrelated variables called principal components. We plotted a variance vs component# graph and we found that the optimal PCA number is 6.



Critical Evaluation: Then, we performed KNN for various K values and plotted accuracy for each KNN. Ultimately, we found that maximum accuracy 77% is achieved for K value = 10

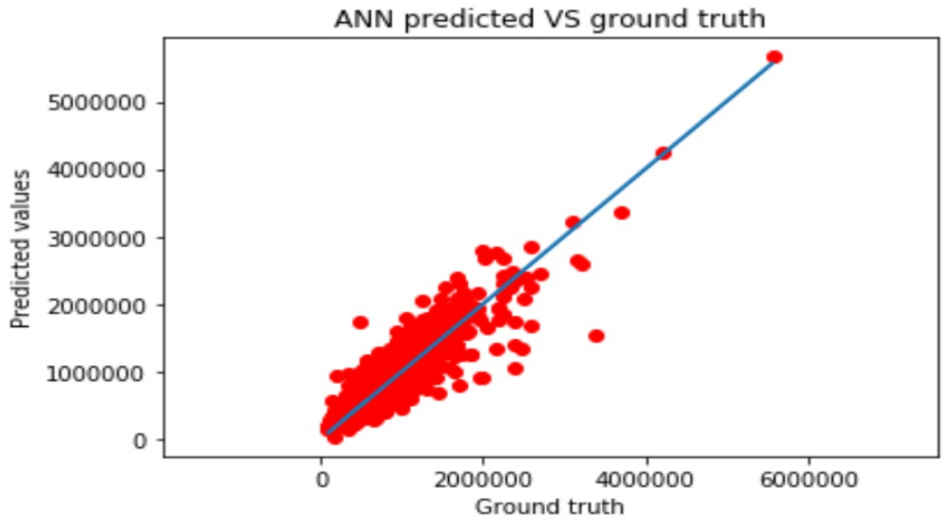


4.4 Artificial Neural Networks

Results: An artificial neural network is based on a collection of connected units or nodes called artificial neurons, which loosely models the neurons in a biological brain. In our model,

1 Input layer - 14 neurons, 1 Output layer - 1 neuron, Epoch: 500, Activation function: rectified linear unit, Keras model: Sequential,

We plotted ANN predicted values against the ground truth. Also, we did performance tuning, and the maximum accuracy achieved is 87%



ANN Performance Tuning

No of layers	Neuron structure	Accuracy
2	64->64->1	75.2
2	50->50->1	75
2	80->80->1	76
2	120->120->1	79
2	130->130->1	87
3	120->120->120->1	86
4	120->120->120->120->1	84

Critical Evaluation: Very less number of neurons result in under-fitting and too many neurons result in over-fitting. Also, too many hidden layers causes over-fitting.

5 Conclusion

5.1 Lessons Learned:

From all the results generated from the above models, we conclude that Random forests and Artificial neural networks is the most efficient models for the given house price prediction data set.

Model Name	Factors	Accuracy
Linear Regression	Standardizing all attributes	70%
KNN without PCA	Standardized with 10NN	79%
KNN with PCA	Standardized with 10NN and 6 Components	77%
Random Forest	Standardized, estimators are 200	88%
ANN	Standardized, 2 hidden layers	87%

References

[1]Taheri, S. and Mammadov, M., 2013. Learning The Naive Bayes Classifier With Optimization Models. In International Journal of Applied Mathematics and Computer Science. pp. 787 – 795.

[2]https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

[3]Dubin Robin A 1998 Predicting Housing Prices using Multiple Listings Data Journal of Real Estate Finance and Economics 17 35-59

[4]https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[5]https://www.stat.berkeley.edu/ breiman/randomforest2001.pdf

[6]https://www.analyticsvidhya.com/blog/2018/08/k-nearest-neighbor-introduction-regression-python/

GitHub

`https://github.com/arjun-0896/Automated-Learning-and-Data-Analysis`