

# Predicting House Price for King County

Link to Demo: <https://youtu.be/KIYB4hx7GMc>

Anirudh Paranjothi – anirudh.paranjothi@ou.edu

Mohammad Mukhtaruzzaman - mukhtar@ou.edu

Aedh Alwadaeen – Aedh.M.Alwadaeen-1@ou.edu

Abinash Borah – abinashborah@ou.edu

## **ABSTRACT**

House price prediction is a ubiquitous challenge for urban life. In this application, we have used Random Forest Regression algorithm to build a model from the data on sale prices of houses in King County, Washington to predict price of a house based on some input criteria. House price depends on some attributes such as number of bedrooms, number of bathrooms, size of the living area, size of the lot area, number of floors etc. We have used historical data for the area for a whole year from May 2014 to May 2015 and used the historical data to build our predictive model. The application will take input from user for their targeted house and will predict the price based on those attributes.

## **Section 1: Introduction**

House price prediction is an important part of the urban life for many reasons and to many groups. When people want to buy a house, as compared to other criteria, price plays the vital role in choosing a house. Identifying the factors which affect house price is a strong economic indicator [Bency, 2017], and an important topic in the economic literature [Montero, 2017]. House price prediction is also termed as a classic problem in econometrics [Osland, 2010]. We can also learn how strong the economy of the region based on their house price and other attributes. So, if Government or any non-government organization wants to develop the area, then they can use this result to choose the area properly. If an area has a lower mean price compared to another area, we can conclude that economic condition of the people of that area is not as good compared to the other areas.

Predicting the price of a house based on certain criteria is a challenging task. It requires a lot of effort to make an accurate prediction. Traditional house price prediction models don't use rigorous data mining tasks to predict the price, consequently their prediction is not accurate. Therefore, a house price prediction model was necessary to fill this gap. Our application will apply Random Forest Regression on data of sale prices of houses along with various characteristics of the houses over a period. Based on several attributes of the houses and sale prices, the application developed a model to predict the price of a house based on different input criteria.

In Section 2, we will discuss related work on this area. Our work with simulation results will be described in detail in Section 3. We will conclude the report with the conclusions and future work in Section 4.

## Section 2: Related Works

Many research papers have been addressed this important issue. [Gan, 2015], [Bency, 2017], [Cebula, 2009], [Lu, 2014] have worked to solve this problem investigating a single algorithm. But house price prediction can vary based on the training data set. Different algorithms perform well for different data sets. For better estimation, we need to apply various algorithms to choose the best one by comparing the results obtained with each of the algorithms. To overcome this problem, we have experimented with four different algorithms and chosen the best one to implement.

In [Krizhevsky, 2012] and [Simonyan, 2015], Deep Convolutional Neural Networks have been used to represent information from large scale data sets. [Khosla, 2014], [Ordonez, 2014] and [Bessinger, 2016] have emphasized the correlation between visual features using the street view imagery. Many papers have relied on satellite images [Mnih, 2010], [Hu, 2007], [Jean, 2016], [Workman, 2015], [Cheng, 2016] and [Meng, 2012] to predict house price. However, these researches highly depend on the images and geographical information to predict the price, which doesn't reflect the practical scenario. They can provide a mean price for a region but can't provide a good estimation for a single house.

Considering all the works that exist in the literature, a predictive model was necessary for housing price prediction, especially for urban areas like King County, Washington. To fill this gap, we performed experiments with four different regression algorithms and implemented the one producing the best results on the data set.

## Section 3: The Proposed Work

Our objective for this project is to develop a predictive model using regression on data of sale prices of houses. Using this model, we have developed an application to predict the house price for a specified area; in this case, King County, Washington. If a potential buyer wants to buy a house in King County, then the application will provide an estimate of price based on the features of the house. In addition, if a house owner from King County wants to sell his/her house then this application will provide an estimate for selling price. Our application will work for not only King County area, this will also work for any area if data set can be obtained with similar attributes.

### 3.1: Dataset and Source

The dataset contains sale prices of houses in King County, Washington, USA. It contains data from May 2014 to May 2015 which comprises of 21613 records and 19 attributes along with house id and house price which is the target attribute. The size of dataset is 2.4MB and it was collected from Kaggle (<https://www.kaggle.com/harlfoxem/housesalesprediction/data>) on Sep 15, 2017. The description of the attributes is given below:

Sl.	Attribute	Data Type	Meaning	Size
1	id	Integer	Unique identifier for each house	2 bytes
2	date	Date	Selling Date of the houses	1 byte
3	price	Integer	Actual selling price	2 bytes
4	bedrooms	Integer	Number of bedrooms	1 byte

5	bathrooms	Float	Number of bathrooms	1 byte
6	sqft_living	Integer	Size of living area	1 byte
7	sqft_lot	Integer	Size of the lot area	1 byte
8	floors	Integer	Number of floors	1 byte
9	waterfront	Integer	Water Front availability	1 byte
10	view	Integer	The view of the unit	1 byte
11	condition	Integer	Conditions of the unit	1 byte
12	grade	Integer	Number of rooms above grade level	1 byte
13	sqft_above	Integer	Size of the area above the ground	1 byte
14	sqft_basement	Integer	Size of the basement	1 byte
15	yr_built	Integer	Original construction year	1 byte
16	yr_renovated	Integer	The year of last renovation	1 byte
17	zipcode	Integer	Zip Code of the unit	1 byte
18	lat	Double	Latitude of the House	8 bytes
19	long	Double	Longitude of the House	8 bytes
20	sqft_living15	Integer	The average living area of 15 closest houses	1 byte
21	sqft_lot15	Integer	The average Lot area of 15 closest houses	1 byte

A sample of the data set is shown below:

id	date	price	bedrooms	bathrooms	sft_living	sft_lot	floors	waterfront	view	condition	grade	sft_above	sft_basement	yr_built	yr_renovated	zipcode	lat	long	sft_living15	sft_lot15
7129300520	20141013T000000	221900	3	1	1180	5650	1	0	0	3	7	1180	0	1955	0	98178	47.5112	-122.257	1340	5650
6414100192	20141209T000000	538000	3	2.25	2570	7242	2	0	0	3	7	2170	400	1951	1991	98125	47.721	-122.319	1690	7639
5631500400	20150225T000000	180000	2	1	770	10000	1	0	0	3	6	770	0	1933	0	98028	47.7379	-122.233	2720	8062
2487200875	20141209T000000	604000	4	3	1960	5000	1	0	0	5	7	1050	910	1965	0	98136	47.5208	-122.393	1360	5000
1954400510	20150218T000000	510000	3	2	1680	8080	1	0	0	3	8	1680	0	1987	0	98074	47.6168	-122.045	1800	7503
7237550310	20140512T000000	1.23E+06	4	4.5	5420	101930	1	0	0	3	11	3890	1530	2001	0	98053	47.6561	-122.005	4760	101930
1321400060	20140627T000000	257500	3	2.25	1715	6819	2	0	0	3	7	1715	0	1995	0	98003	47.3097	-122.327	2238	6819
2008000270	20150115T000000	291850	3	1.5	1060	9711	1	0	0	3	7	1060	0	1963	0	98198	47.4095	-122.315	1650	9711

### 3.2: Architecture

The application has a GUI wherein a user can input different attributes of a house for predicting the price. On the backend, a regression model built from the data set using the Random Forest Regression algorithm will take the inputs and predict a price. This price will be displayed on the GUI of the application.

We have used R (<https://www.r-project.org/>) as our programming language, R Studio (<https://www.rstudio.com/>) as Integrated Development Environment (IDE) and Shiny (<https://shiny.rstudio.com/>) for development of GUI.

### 3.3: Algorithm

We are using regression as our data mining task; moreover, we have used anomaly detection to remove outliers. Regression algorithms estimate the value, i.e., price, as a function of the predictors in the training data set. These relationships between predictors and price are summarized in a model, which can then be applied on different input data in which the price is unknown.

We experimented with Random Forest [Biau, 2012], [Breiman, 2001], decision tree CART [Nilima, 2012], Bagging CART with Bootstrap aggregation [Breiman, 1996], and Multiple Linear Regression [Simon, 2003], [Uyanic, 2013]. While working with regression algorithms [Gupta, 2015], it was found that few regression techniques work only if the target attribute is nominal, while few others work if the target attribute is binary. However, all the mentioned four regression algorithms are capable to predict housing price based on several attributes. Among them, performance of the Random Forest on our data set was better in comparison to the others.

**Random Forest Algorithm:** Random Forest works based on regression and determines whether any significant relationship is present or not between the target variable  $Y$  and its predictor variables ( $X$ ). They are a variation of decision trees by reducing the attributes available to build a tree at each decision point. They construct a multitude of decision trees at training time and give a mean prediction as output of the individual trees. The generalization error for forests converges as the number of trees in the forest becomes large. The generalization error of random forests depends on the strength of the individual trees in the forest and the correlation between them. Random forests minimize the habit of overfitting for the training data set.

For random forests, if we have  $k$  number of trees such that  $1, 2, \dots, k$ ; then for the  $k$ th tree, a random vector  $\Theta_k$  is generated, which is independent of the past random vectors  $\Theta_1, \Theta_2, \dots, \Theta_{k-1}$  but distribution is the same. If  $x$  is an input vector, then a tree is grown using the training set and  $\Theta_k$ , which generates a resulting classifier  $h(x, \Theta_k)$ . Here,  $\Theta$  consists of several independent random integers between 1 and  $K$ .



### 3.4 Data Preprocessing

We analyzed the data set and performed few preprocessing works that was necessary. There were no missing values in the data set. However, there were outliers in the attributes, viz., price, sqft\_lot, sqft\_living and sqft\_above. Initially we had 21,613 tuples in our data set. We removed 3,490 tuples as they contained outliers. After removing the outliers, we have 18,123 instances. We have removed few attributes from our data set, viz., id, date, yr\_renovated, sqft\_basement, yr\_renovated, zipcode, lat, long, sqft\_living15, sqft\_lot15 for prediction purpose since these attributes have no impact on predicting sale prices.

### 3.5 Performance Evaluation and Development Choice:

We used Root Mean Square Error (RMSE) to evaluate the performance of the four regression algorithms on our dataset. The average RMSE values obtained for the different algorithms on the dataset with 10-fold cross validation are summarized below:

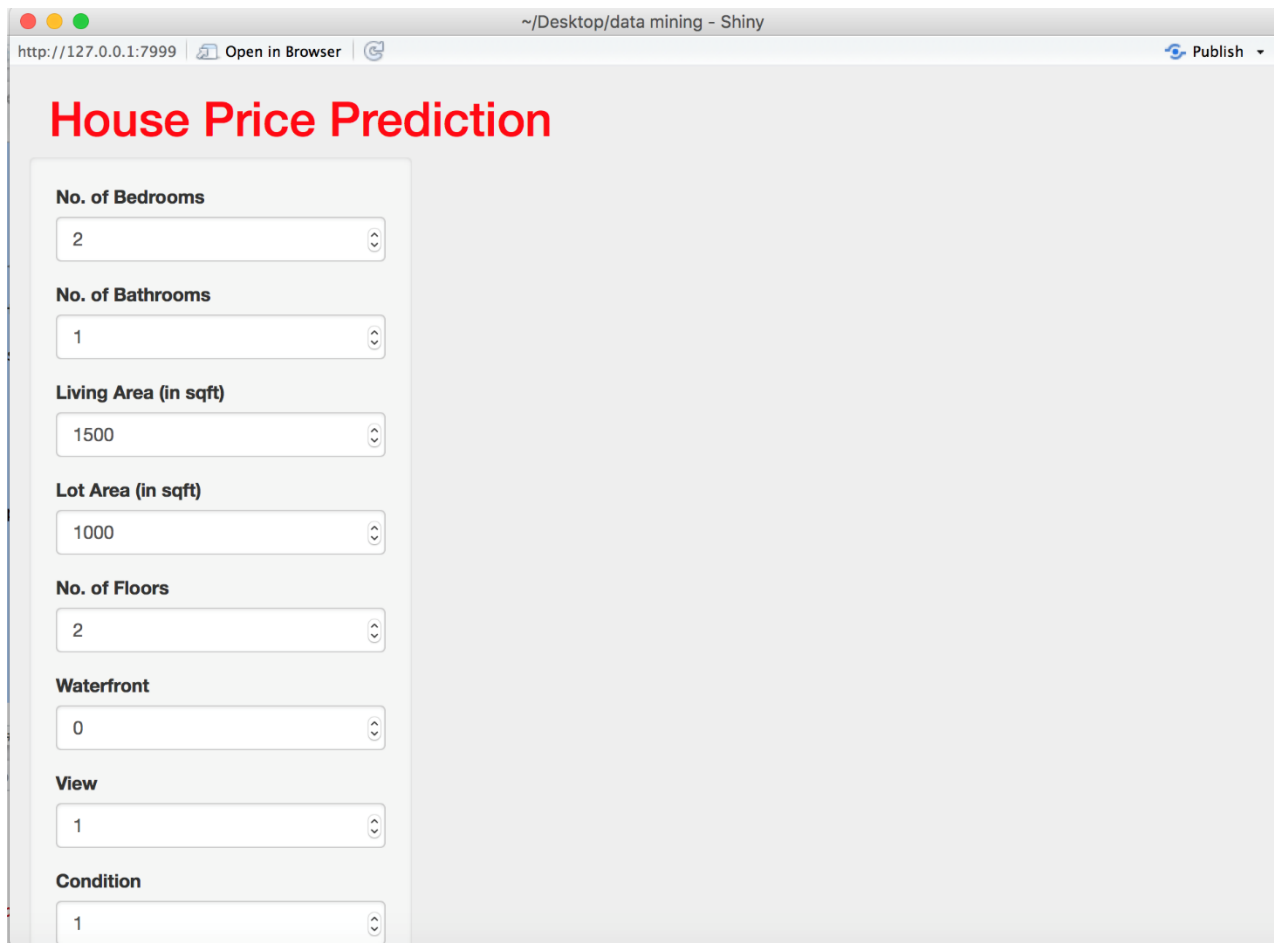
Sl. No.	Algorithm	RMSE
1	Random Forest Regression	125320.9
2	CART Regression	143384.1
3	CART with Bagging	139405.6
4	Multiple Linear Regression	133124.7

Performance of Random Forest algorithm was found to be the best among these algorithms and we decided to implement this algorithm from scratch for using in the application.

### 3.6 How to Use the Application

The application provides textboxes for inputting 11 different attributes for a house. A user needs to input these attributes for a house to predict a price. Based on the user inputs, the application gives a prediction for the price using the model built from the data set using the Random Forest algorithm.

Snapshots of the GUI of the application are shown below.



The screenshot shows a web application titled "House Price Prediction" running in a browser window. The browser's address bar shows "http://127.0.0.1:7999" and the page title is "~/Desktop/data mining - Shiny". A "Publish" button is visible in the top right corner. The application interface features a sidebar with several input fields, each with a label and a numeric value:

- No. of Bedrooms:** 2
- No. of Bathrooms:** 1
- Living Area (in sqft):** 1500
- Lot Area (in sqft):** 1000
- No. of Floors:** 2
- Waterfront:** 0
- View:** 1
- Condition:** 1

The image shows a web form for house price prediction. It has a light gray background. On the left side, there are four input fields stacked vertically. The first field is labeled 'Grade' and contains the value '2'. The second field is labeled 'Sqft Above' and contains the value '1300'. The third field is labeled 'Year Built' and contains the value '2006'. The fourth field is a text input containing the value '458797'. To the right of these fields is a large, empty gray rectangular area.

## Section 4: Conclusion and Future Work

In this work, we have developed an application to predict the house price using Random Forest algorithm. We have built our application for King County. However, we can extend our scope by obtaining data set for the whole state, and possibly for countrywide, if proper data set is available. In future work we can incorporate geographical location as an attribute, because, in real-world, prices vary based on location. Consequently, the prediction will be more accurate.

\*\*\*\*\*

## References:

- [Bency, 2017] Bency A.J., Rallapalli S., Ganti R.K., Mudhakar S., and Manjunath B. S.: “Beyond Spatial Auto-Regressive Models: Predicting Housing Prices with Satellite Imagery”, IEEE Winter Conference on Applications of Computer Vision, pp. 320-329, 2017.
- [Bessinger, 2016] Bessinger Z. and Jacobs N. “Quantifying Curb Appeal”, IEEE International Conference on Image Processing, pp. 4388–4392, 2016
- [Biau, 2012] Biau G., "Analysis of a random forests model", The Journal of Machine Learning Research, Vol. 13, Issue 1, pp. 1063-1095, 2012.
- [Breiman, 1996] Breiman L., "Bagging Predictors", Journal of the American Statistical Association, Vol. 24, Issue 1, pp. 123–140, 1996.
- [Breiman, 2001] Breiman L., “Random Forests”, Statistical Department, University of California, <https://www.stat.berkeley.edu/~breiman/randomforest2001.pdf>, Accessed on Oct 17, 2017.
- [Cebula, 2009] Cebula R.J., “The Hedonic Pricing Model Applied to the Housing Market of the City of Savannah and its Savannah Historic Landmark District”, The Review of Regional Studies, Vol. 39, Issue 1, pp. 9–22, 2009.
- [Cheng, 2016] Cheng G., and Han J., “A Survey on Object Detection in Optical Remote Sensing Images”, International Journal of Photogrammetry and Remote Sensing, pp. 11–28, 2016.
- [Gan, 2015] Gan V., Agarwal V., Kim B., "Data Mining Analysis and Predictions of Real Estate Prices", International Conference on Issues in Information Systems, Vol. 16, Issue 4, pp. 30-36, 2015.
- [Gupta, 2015] Gupta S., "A Regression Modeling Technique on Data Mining", International Journal of Computer Applications, Vol. 116, Issue 9, pp. 27-29, 2015.

- [Hu, 2007] Hu J., Razdan A., Femiani J. C., Cui M., and Wonka P., “Road Network Extraction and Intersection Detection from Aerial Images by Tracking Road Footprints”, IEEE Transactions on Geoscience and Remote Sensing, pp. 4144–4157, 2007.
- [Jean, 2016] Jean N., Burke M., Xie M., Davis W. M., Lobell D. B., and Ermon S., “Combining Satellite Imagery and Machine Learning to Predict Poverty”, Journal of Science, pp. 790–794, 2016.
- [Khosla, 2014] Khosla A., An B., Lim J. J., and Torralba A., “Looking Beyond the Visible Scene”, IEEE Conference on Computer Vision and Pattern Recognition, pp. 3710-3717, June 2014.
- [Krizhevsky, 2012] Krizhevsky A., Sutskever I., and Hinton G. E., “Imagenet Classification with Deep Convolutional Neural Networks”, Journal of Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.
- [Lu, 2014] Lu B, Charlton M, and Fotheringham A.S., “Geographically Weighted Regression with a Non-Euclidean Distance Metric: A Case Study Using Hedonic House Price Data”, International Journal of Geographical Information Science, Vol. 28, Issue 4, pp.660–681, April 28, 2014.
- [Meng, 2012] Meng L., and Kerekes J. P., “Object Tracking Using High Resolution Satellite Imagery”, IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, pp. 146– 152, 2012.
- [Mnih, 2010] Mnih V., and Hinton G. E., “Learning to Detect Roads in High Resolution Aerial Images”, In the Proceedings of the Conference on Computer Vision, pp. 210–223, 2010.

- [Montero, 2017] Montero J.M., Minguez R., and Aviles G.F.: “Housing Price Prediction: Parametric versus Semiparametric Spatial Hedonic Models”, *Journal of Geographical Systems*, Springer, pp. 1-29, 2017.
- [Ordonez, 2014] Ordonez V., and Berg T. L., “Learning High-level Judgments of Urban Perception”, *Proceedings on Conference on Computer Vision*, pp 494–510, 2014.
- [Osland, 2010] Osland L., “An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling”, *Journal of Real Estate Research*, vol. 32, issue 3, pp.289-320, 2010.
- [Patil, 2012] Patil N., Rekha L., Vidya C., “Comparison of C5.0 & CART Classification algorithms using pruning technique”, *International Journal of Engineering Research & Technology*, Vol. 1, Issue 4, pp. 1-5, 2012.
- [Simon, 2003] Simon G., "Multiple Regression Basics", New York University, Stern School of Business, <http://people.stern.nyu.edu/wgreene/Statistics/MultipleRegressionBasicsCollection.pdf> Accessed on Oct 17, 2017.
- [Simonyan, 2015] Simonyan K. and Zisserman A., “Very Deep Convolutional Networks for Large-scale Image Recognition”, *International Conference on Learning Representations*, pp. 1-14, 2015.
- [Workman, 2015] Workman S., Souvenir R., and Jacobs N., “Wide-area Image Geolocalization with Aerial Reference Imagery”, In the *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1–9, 2015.
- [Uyanik, 2013] Uyanik G. K., Guler N., "A Study on Multiple Linear Regression Analysis", *International Conference on New Horizons in Education*, Vol. 106, pp. 234-240, 2013.