Seattle, Kings County House Prediction

Xingrou Mei
Fordham University
Dept. of Computer & Info. Science
Xmei4@fordham.edu

***Abstract*— When people are interested in purchasing or selling a house, the price is the first concern. Although there are many aspects that can predict the house value, the most important factor is the house properties. Understanding the worth of a house based on its properties is essential to price determination. The goal of this paper is to discover which features are strongly correlated with price. This paper implements different regression algorithms like Linear Regression, Bagging, and Random Forest to estimate the accuracy of price prediction. The result states the Random Forest has the best result. Linear Regression performs poorly on this paper. One linear line cannot fit the model accurately.**

Keywords—House prediction, Seattle Real Estate, Data Mining, Regression Task

I. INTRODUCTION

Using machine learning algorithms for house price prediction is commonly researched among scholars. Many studies are focusing on algorithm comparison and trying to find the best algorithm to predict the housing price. For instance, using Linear Regression, Random Forest, and Gradient Boosting on different testing set sizes [6]. Besides using regression techniques, some studies use artificial neural networks to predict the real estate price [7]. Before applying algorithms, it is essential to discover the pattern between price and physical features of a house like geographic area, size of living space, house age, number of bathrooms, etc.

Back in 2010, the housing market in Seattle was not as high as today's market. Between 2010 to 2017, Seattle housing prices skyrocketed by about 83.4% [1]. One of the reasons why it rapidly increases was due to the Amazon Effect, according to many articles and media. The tech giant, Amazon, came with 45,000 jobs when they decided to build their second headquarter in Seattle, Kings County [2,3]. Due to the growth of employment and high-paying job, it increased the housing demand near the headquarter of Amazon [2]. Amazon launched its first Seattle campus in 2015. The dataset for this paper is the sale of houses in 2014 and 2015 in Kings County, Seattle. It will be interesting to see if the price is related to the Amazon Effect. However, the dataset only includes the characteristics of houses and their value. It might be hard to estimate the correlation between Amazon and the price. But apply the coordinates may help to discover if the housing price near Amazon Headquarters is relatively high in Seattle or not. It will be an interesting question alongside the research.

To conclude, this paper aims to focus on finding the correlation between features and price and predict the value of a house based on its features. This paper will use Linear Regression, Decision Tree, and Random Forest algorithms to predict the accuracy of house values. 10-fold cross-validation is for avoiding overfitting the model. The metrics that will determine the performance of the algorithms will be Root Mean Square Error (RMSE), Mean Square Error (MSE), and Relative Absolute Error (RAE).

II. EXPERIMENT METHODOLOGY

A. Dataset

The dataset contains house sale prices in Seattle, Kings County, between 2014 and 2015. It includes 2 features and 21,613 examples [10].

In the bathroom column, there are decimal numbers instead of whole numbers. In the United States, there are different types of bathrooms: full, master, three-quarter, half, and quarter bathrooms. Full bathroom has a toilet, sink, shower, and bathtub/shower, and it stands for 1. Three-Quarter bathroom has a toilet, sink, and a separate bathtub/shower, 0.25. Half-bathroom has a toilet and a sink, 0.5. The Master bathroom is a bathroom inside the master bedroom with either a full or three-quarter bathroom [5]. For example, 1.75 indicates 1 full, 1 half, and 1 three-quarter bathroom in a house. There is only one binary feature in the dataset, waterfront.

Table 1 includes detailed information of the dataset with the mean, minimum, and maximum.

Table 1: Dataset	**House Feature Detail information**			
	Description	***Average***	***Min value***	***Max value***
bathrooms	number of bathrooms	2.12	0	8
bedrooms	number of bedrooms	3.37	0	33
condition	condition of house rated in the range 1 to 5	3.41	1	5
year	2014-2015	-	2014	2015
floors	number of floors	1.49	1	3.5
grade	grade of house ranging from 1 to 13	7.65	1	13
id	unique identification for each house	-	-	-
lat	latitude position of the house	-	47.15590	47.77760
long	longitude position of the house	-	-122.5190	-121.3150
price	price of house in US dollars	540,292	78,000	7,700,000
sqft_above	the surface area of house in square feet above ground level	1,788.63	370	9410
sqft_basement	the surface area of house in square feet below ground level	291.71	0	4820
sqft_living	square foot of living area	2,080.34	370	13,540
sqft_living15	average house square footage of the 15 closest houses	1,986.65	399	871,200
sqft_lot	square foot of lot	15,099.83	520	1,651,359
sqft_lot15	average lot square footage of the 15 closest houses	12,758.66	651	871,200
view	rating of view rated 0 to 5	0.23	0	4
waterfront	house with a waterfront view binary: 1 or 0	-	0	1
yr_built	the year of house constructed	-	1900	2015
yr_renovated	the year of house renovated	-	0	2015
zipcode	5-digit zip code	-	-	-

B. *Data Processing*

In the dataset, there are some contradicted examples. There is a house with 33 bedrooms, but only 1620 square feet of living area. It does not seem to make sense, and it was removed from the dataset. There are some houses with 0 bathrooms and 0 bedrooms, and it was removed as well.

- **Feature selection:**

The data processing part is obtained by Python with Matplotlib and Seaborn. The fastest way to find the relationship between features is by using correlation Metrix. In this paper, we mostly focus on features that are correlated with price. Below **Table 2** shows all the correlations with the price before data processing and after data processing. It is sorted with the highest to lowest correlation.

Table 2: Correlation	**Price vs. Features**	
	Corr. before processing	***Corr. after processing***
sqft_living	0.70204	0.70193
grade	0.66746	0.66796
sqft_above	0.60557	0.60539
sqft_living15	0.58537	0.58527
bathrooms	0.52513	0.52592
view	0.39735	0.39738
sqft_basement	0.32384	0.32379
bedrooms	0.30834	0.31596
lat	0.30692	0.30669
waterfront	0.26633	0.2664
floors	0.25679	0.25682
yr_renovated	0.12644	Removed
sqft_lot	0.08966	0.08988
sqft_lot15	0.08246	0.08285
yr_built	0.05398	Removed
condition	0.03639	0.03603
long	0.02157	0.02205
year	0.00355	0.00374
yr_old	-----------	-0.105631
zip_code	-0.05340	Removed
id	Removed	

Feature 'id' is a unique number for each sample; thus, it is an irrelevant feature. This paper uses the filter technique to remove any redundant features that are not helpful. **Table 2** above shows that the correlation between the features 'zip_code,' 'year,' 'long,' 'condition,' 'yr_built,' 'sqft_lot,' sqft_lot15,' and 'yr_renovated' with 'price' are very low. However, some features might be useful after some adjustments. For instance, 'yr_built,' 'yr_renovated,' and 'year' may be able to calculate the age of a house. To find a location specifically, 'long' is necessary with 'lat.' 'condition' might be related to 'grade.' Although 'zip_code' can identify a location of an area, it is not as specific as 'long,' and 'lat.' Therefore, 'zip_code' is removed. Based on **Table 2,** 'sqft_living' has the highest correlation with price. It indicates that the larger the living space of a house, the higher the price. The second highest is 'grade,' the better the rating, the higher the price. 'sqft_above' is the third highest, but the correlation between 'sqft_living' and 'sqft_above' has a high correlation, 0.88. It shows that both of them are somewhat identical to each other. These two features are very similar to one another, but not sure if 88% of similarity is good enough to keep one only. Therefore, none of them are removed.

- **Feature Creation:**

Since the correlation between price with 'yr_built' and 'yr_renovated' is relatively low, then it might be better to check if the age of the house. A new column, called “yr_old,” was added to the dataset, and it is the age of the house. To calculate the age, the number of 'yr_renovated' minus the number of 'yr_built' or 'year' minus 'yr_built.' The new column, 'yr_old,' shows that they are still not highly correlated with price. However, the scatter plot, **Figure 1**, shows that houses with lower age, meaning newer, have higher price over about five million dollars and older houses do not. 'yr_old' might be useful for analysis, and only 'yr_built' and 'yr_renovated' were removed.

Waterfront is a binary feature and unbalanced with a large number of houses that do not have a waterfront. The box plot, **Figure 2**, shows the relationship with price. It shows that most of the houses that have a waterfront view have a higher value in price.

Figure 1: House age vs. Price

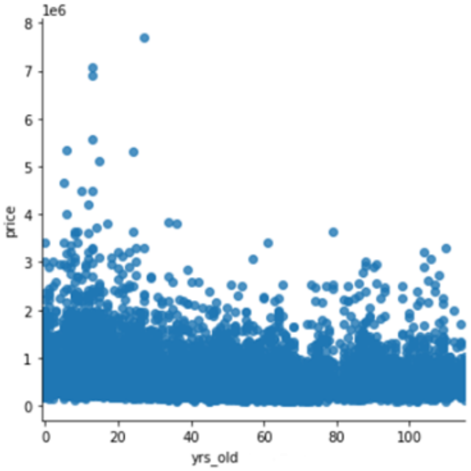
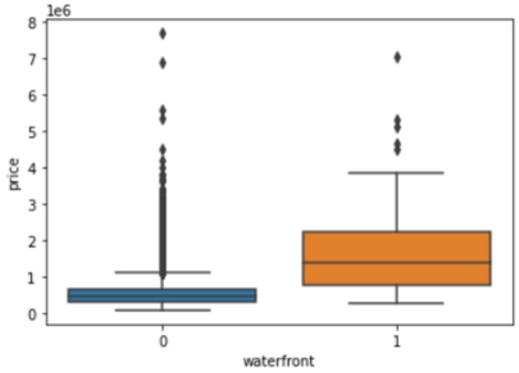
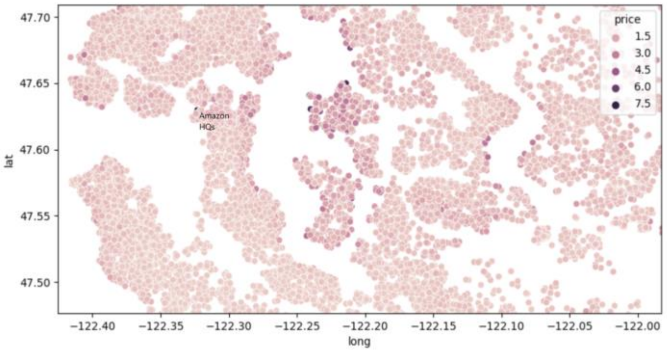


Figure 2: Waterfront vs Price



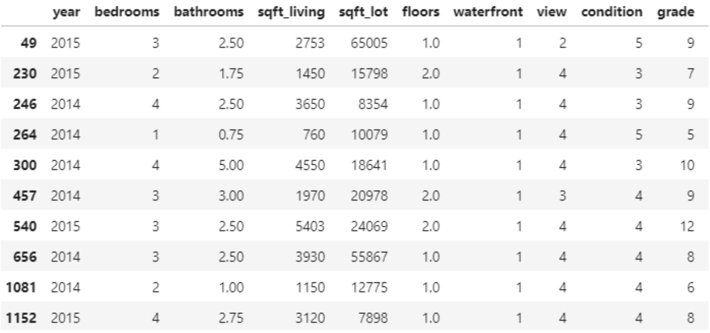
Using a scatter plot with an x-axis of longitude and a y-axis of latitude can easily show a map-like chart. Figure 3 shows all the houses in Seattle based on housing price, and the higher the price, the darker the color in the plot.

Figure 3: Kings County, Seattle House Sales



Amazon HQs Coordinates:
Amazon Day 1: 47.615868, -122.339850
Amazon Doppler: 47.615144 N 122.338578 W

Chart 1: Waterfront vs. View



	year	bedrooms	bathrooms	sqft_living	sqft_lot	floors	waterfront	view	condition	grade
49	2015	3	2.50	2753	65005	1.0	1	2	5	9
230	2015	2	1.75	1450	15798	2.0	1	4	3	7
246	2014	4	2.50	3650	8354	1.0	1	4	3	9
264	2014	1	0.75	760	10079	1.0	1	4	5	5
300	2014	4	5.00	4550	18641	1.0	1	4	3	10
457	2014	3	3.00	1970	20978	2.0	1	3	4	9
540	2015	3	2.50	5403	24069	2.0	1	4	4	12
656	2014	3	2.50	3930	55867	1.0	1	4	4	8
1081	2014	2	1.00	1150	12775	1.0	1	4	4	6
1152	2015	4	2.75	3120	7898	1.0	1	4	4	8

Waterfront is not correlated with price, but the house with a waterfront mostly have a view rating of 4.

C. *Algorithms*

Linear Regression is basic and simple analysis, and it uses the mathematical relationship between variables or features by fitting one linear equation.

Ensembles Regressions:

The ensemble is a method that uses a group of classifiers to make one prediction together.

Bagging is a technique that creates many random samples in the dataset with replacement, then it builds a classifier on each bootstrap sample, and each case has a probability of (1-1/n)^n of being picked. Lastly, a decision is combined with majority voting.

Random Forest is a popular ensemble of decision trees. It is similar to the Decision Tree Regression, but the final decision is finalized with majority voting. When splitting an internal node, the best split is chosen randomly from a subset of features. Random Forest is an ensemble method that uses bagging and decision trees.

D. *Metrics*

There are some measurements to evaluate the performance of the algorithms.

- Root Mean Square Error (RMSE)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(P_i - \hat{P}_i)^2} \tag{1}$$

- Mean Square Error (MSE)

$$MAE = \frac{1}{n}\sum_{i=1}|P_i - \hat{P}_i| \tag{2}$$

- Relative Absolute Error (RAE)

$$RAE = \frac{\sum_{i=1}^{n}|p_i - a_i|}{\sum_{i=1}^{n}|\overline{a} - a_i|} \tag{3}$$

Root Mean Squared Error is the square root of the average of squared differences between the predicted value and actual value. It measures the average magnitude of the errors, meaning it measures how far are the data points are to the regression line on average [9].

Mean Absolute Error measures the average of the difference between the predicted value and actual value [9].

Relative Error measures the difference between the absolute error and the actual value by percentage.

III. RESULT

This study implemented three regression algorithms in the Seattle House Sale dataset, Linear Regression, Bagging, and Random Forest. Two sections operated in the dataset, 10-fold cross-validation, and training 66% of the dataset and test 33% of the dataset. **Table 3** shows the result for all the algorithms with 10-fold cross-validation and Train/Test split.

The results were achieved by WEKA and it is reported in Table 3. There are two different accuracies, one using 10-fold cross-validation and one splitting training set and test set to 66% and 33% with respectively. The accuracy for ensemble methods, Bagging and Random Forest, has a better result with cross-validation than the training test set. In Linear Regression, the training test set performance better than the 10-fold cross-validation. Metrics are displayed in Table 4, MAE, RMSE, and RAE. Based on the metrics, the algorithm with the lowest number is Random Forest. The best metric to determine the performance of the regression model is RMSE. This error shows the difference between the data points and the regression line on average. It is directly connected to the regression model, and it indicates how well the model is fitting. Random Forest regression has the lowest error of $131,063. It seems like a larger number, but for this dataset, it is considered a small number. The minimum price for all the houses in this dataset is $78,000, and the average price is $540,292. Bagging is very similar to Random Forest with a $3614 difference. For MAE, the difference between the predicted price and the actual price is $70,779 on average. RAE is similar to MAE but by percentage. Linear Regression is a 56.28% error rate between the predicted price and the actual price. It is a very high error rate of more than 50%. Overall, Linear Regression has the highest number and percentage among all.

Table 3	**Algorithms Comparison**	
	10-fold cross validation	***0.66/0.33 Train/Test***
Linear Regression	0.8218	0.8256
Bagging	0.9268	0.9162
Random Forest	0.9354	0.9167

Table 4	**Metrics**		
	MAE	***RMSE***	***RAE***
Linear Regression	133792.98	197387.60	56.28 %
Bagging	74393.04	138540.78	31.79%
Random Forest	70779.19	131063.41	30.24%

CONCLUSION

To conclude this paper, the best algorithm is Random forest with an accuracy of 91.62%, the lowest error rate of 30.24%. Most importantly, it has the lowest RMSE among all other regression models. This means the model fits better than the others. It is a relatively good result because, before the prediction, I expected the percentage within 10%. Since the dataset have many features that are not highly correlated with price, and only 5 features are higher than 50%. Based on Table 2, the highest correlation with the price is sqft_living. This indicates that the bigger the living area, the higher the price. The grade of a house is also very important to the price with a 66% impact on high prices.

Although houses with a waterfront view are not highly correlated with the price due to imbalance data, Figure 2 shows that house with a waterfront view is much higher than the houses without it. Also, the view is somewhat correlated with the price with a 40% correlation. Most houses with a view with a rating of 4 have a waterfront, shown in **Chart 1**. This is indicating that usually houses with a waterfront has a very nice view. Therefore, a waterfront view also affects the value of the house.

For the Amazon effect, Amazon is located in the downtown of Seattle, in the coordinate 47.60-47.65 and -122.30. In Table 3, the majority of the expensive houses are in the area called Bellevue. In this research, most of the expensive houses in this dataset are not around Amazon headquarter. However, based on this one dataset cannot give a conclusion that houses near Amazon are not expensive. Places in downtown could be mostly apartments, and this dataset only contains houses.

REFERENCES

[1] Hanley Wood Data Studio, "The Amazon Effect: Housing Affordability," Builder, 02-Apr-2018. [Online]. Available: https://www.builderonline.com/money/economics/the-amazon-effect_o#:~:text=The "Amazon Effect" will drive,price of over $1 million.

[2] Ewing & Clark Inc., "How Amazon Has Impacted Seattles Housing Market," Ewing & Clark Inc., 14-Sep-2020. [Online]. Available: https://www.ewingandclark.com/how-amazon-has-impacted-seattles-housing-market-2/.

[3] "The Amazon effect: The good, the bad and the ugly — straight from Seattle," The Business Journals. [Online]. Available: https://www.bizjournals.com/washington/news/2018/07/20/the-amazon-effect-the-good-the-bad-and-the-ugly.html.

[4] J. Reynolds and A. Details, "10 Year Chart Of The Seattle Real Estate Market Is Mind Blowing, Up 93% Since The Bottom," UrbanCondoSpaces, 17-Jan-2018. [Online]. Available: https://www.urbancondospaces.com/10-year-chart-seattle-real-estate-is-mind-blowing/.

[5] L. Wallender, "5 Types of Bathrooms," The Spruce. [Online]. Available: https://www.thespruce.com/types-of-bathrooms-4800093.

[6] Uzut, Gulsum & Buyrukoglu, Selim. (2020). Prediction of real estate prices with data mining algorithms. Euroasia Journal of Mathematics Engineering Natural and Medical Sciences. 7. 77-84.

[7] C. Xiaochen, W. Lai, X. Jiaxin, "House Price Prediction Using LSTM," presented at The Hong Kong University of Science and Technology, Hong Kong.

[8] J. Demmitt, "Amazon launches new era with opening of first tower at new Seattle campus," *GeekWire*, 14-Dec-2015. [Online]. Available: https://www.geekwire.com/2015/amazon-launches-new-era-with-opening-first-tower-at-new-seattle-campus/.

[9] Jj, "MAE and RMSE - Which Metric is Better?," Medium, 23-Mar-2016. [Online]. Available: https://medium.com/human-in-a-machine-world/mae-and-rmse-which-metric-is-better-e60ac3bde13d.

[10] Harlfoxem, "House Sales in King County, USA," Kaggle, 25-Aug-2016. [Online].Available: https://www.kaggle.com/harlfoxem/housesalesprediction.