What are the features that influence housing price in King County? The best predictive model.

PHAN THI THU TRANG

Code Link: https://github.com/TiaPhan/Housing-Price

Highlights

- Houses near waterfront, Medina, Clyde Hill and Mercer Island, with big living area and new condition are more expensive than the other.
- Housing price is stable during the year. April and May have the highest sales volume.
- Extreme Gradient Boost model is more effective compared to other machine learning predictive models.

Abstract

This paper proposes and tests an integrated model to explain how size, location, number of bathroom and bedroom, floor, condition, year built, year renovated influence housing price in King County of USA. In addition, this paper investigates the effectiveness of using different machine learning models in predicting future housing price and put them in comparison to find the best predictive model. A study (Study 1) was conducted to examine the mechanism underlying the impact of house size, features, floor, condition, year built, year renovated on King County's house price. The results indicate that the price for new houses near Lake Washington and Lake Sammamish, Medina, Clyde Hill and Mercer Island with living areas bigger than 7000 square feet are more costly. A follow-up experimental study (Study 2) was performed to compare the effectiveness of different machine learning predictive models in predicting future housing price. To explore various impacts of features on predictive methods, this paper will apply both traditional and advanced machine learning approaches to investigate the difference among several advanced models. The results showed that Extreme Gradient Boost is the most effective model among Lasso Regression and Random Forest.

Keywords: House Price Prediction, Machine Learning.

1. Introduction

The trend of house prices is always a controversial topic as its fluctuation will pose a huge effect on the entire economy. A rise in housing price can have large effects on consumption. First, if the value of a house increases, this may encourage homeowners to consume more because they believe that their wealth has increased. Second, homeowners would be more willing to borrow because the house can be used as collateral, reducing credit constraints. When there is a long-term change in housing prices, resulting changes in consumption can be approximately, using homeowners' marginal propensity to consume and their home values (David Berger, Veronica Guerrieri, Guido Lorenzoni, Joseph Vavra, 2017). A rise in housing price can also redistribute wealth within an economy – increasing the wealth of homeowners, but reducing effective living standards for those who do not own a house. On the other hand, a sharp drop in housing price can contribute to economic recession since it adversely affects consumer confidence, construction and leads to lower economic growth. Housing is an essential sector of the economy but also one that has been the source of vulnerabilities and crises. Hence, there is a great need to know what impacts housing price and how to predict house price in the future.

In recent years, due to the growing trends in Big Data, how to mine valuable data from enormous data set has been a big challenge. Implementing machine learning algorithms in predicting has become a vital approach due to its accuracy. They are used every day to make critical decisions in medical diagnosis, stock trading, energy load forecasting, and more. This paper utilizes machine learning algorithms to achieves housing price prediction in King County, USA. I used the "House Sales in King County, USA" dataset uploaded to Kaggle by harlfoxem. By applying several methods on this dataset, we could validate the performance of each individual approach. The highest accuracy is 88.5% on the test set, which belongs to the Extreme Gradient Boost model.

The rest of the paper is structured as follows: Section 2 for literature review, Section 3 illustrates the details of the methodology, Section 4 conducts study 1, Section 5 conducts study 2, Section 6 compares and selects the best model, Section 7 conclude the paper with remarks, Section 8 mentions about future work to improve the model.

2. Literature Review

Before this paper, there were already so many projects about real estate. According to Kusan et al. (2010), main factors which affect housing price can be classified into three types: house factors, environmental factors, and transportation factors. When people consider purchasing a house for living purposes, these factors are the main determinants for living quality. Buyers with family members would likely focus more on the essential features of the house such as house size, number of rooms, which have a significant impact on living quality. For the predictive model, there have been a large number of empirical studies analyzing land prices over the last two decades. Mundy and Kilpatrick (2000) showed the usefulness of time-series regression model which used economic data to provide forecast of land price in moving market. Wilson et al (2002) studied the residential property market accounts for a substantial proportion of UK economic activity. Wang and Tian (2005) used the Artificial Neural Networks to forecast the future trend of housing market.

3. Methodology

3.1. Data description

“House Sales in King County, USA” is a dataset with 21,613 data with 21 variables representing housing prices traded between May 2014 and May 2015. These variables, which served as features of the dataset, were then used to predict the price of each house in King County.

Among the 21 features, eight of them are the continuous numerical variables, that describe the area dimensions in measurements and the geographical location of the house. These continuous variables provide a basic view of the overall structure and information of the house. The rest of the attributes are discrete variables, which provide some more detailed information on components of the house. Most of them quantify the number of items in the house, for instance, the number of bedrooms, bathrooms, waterfront, and floor. Some others indicate the background of the house, such as year of building, year of innovation and previous selling price and date.

Table 1. Description of the attributes

Attribute Name	Data Type	Description

id	float64	Unique ID for each home sold
date	numeric	Date house was sold
price	float64	Price of each home sold
bedrooms	int64	Number of bedrooms in a house
bathrooms	float64	Number of bathrooms in a house
sqft_living	int64	Square footage of a house
sqft_lot	int64	Square footage of the land space
floors	float64	Total floors in a house
waterfront	int64	House which has a view to a waterfront
view	int64	House has been viewed
condition	int64	How good the condition is overall
grade	int64	Overall grade given to the housing unit, based on King County grading system
sqft_above	int64	Square footage of the interior housing space that is above ground level
sqft_basement	int64	Square footage of the basement
yr_built	int64	The year the house was built
yr_renovated	int64	The year when the house was renovated
zipcode		Zip code of the area
lat	float64	Latitude coordinate
long	float64	Longitude coordinate
sqft_living15	int64	The square footage of interior housing living space for the nearest 15 neighbors

sqft_lot15	int64	The square footage of the land lots of the nearest 15 neighbors

3.2. Data Preprocessing

All the variable in the dataset don’t have any missing data. Below are a few features engineering processes which were done to cleanse the dataset:

- Remove id, zip code due to their unimportance.
- Convert bathrooms into integer.
- Create ‘age’ column to understand the age of a house.
- Create ‘renovated’ column, setting 0 to has renovated, 1 to hasn’t renovated.

4. Study 1

4.1. Measurement

This study investigated the impacts of all the features on housing price to see which features influence price the most. Figure 1 demonstrates that most data concentrated in the West of King County, especially in Seattle. Data are rare in cities located in the East of the county, such as Snoqualmie and Skykomish. The most expensive houses centralize near the Lake Washington and Lake Sammamish, also houses in the Medina, Clyde Hill and Mercer Island areas are more expensive than the other.

Besides the location features, other features of the house also significantly contribute to the rise in prices. In Figure 2, 3, 4 we can see that newer house with higher grade and bigger house size lead to higher housing price. Figure 5 and 6 shows us prices are pretty stable and most of the houses were sold in April and May.

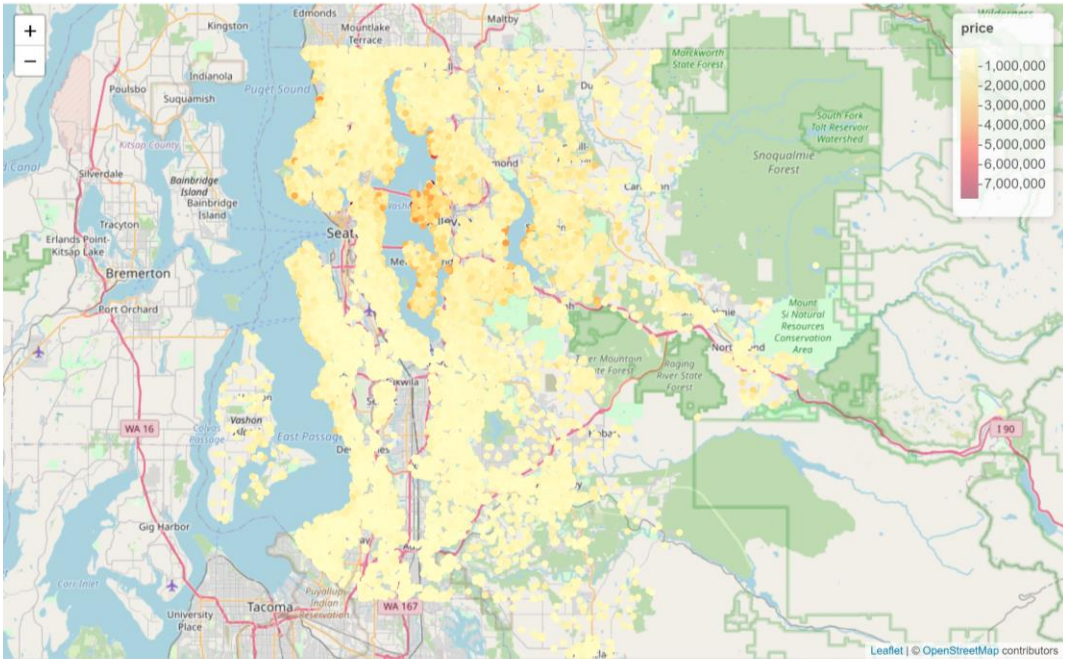

Figure 1. Price Distribution

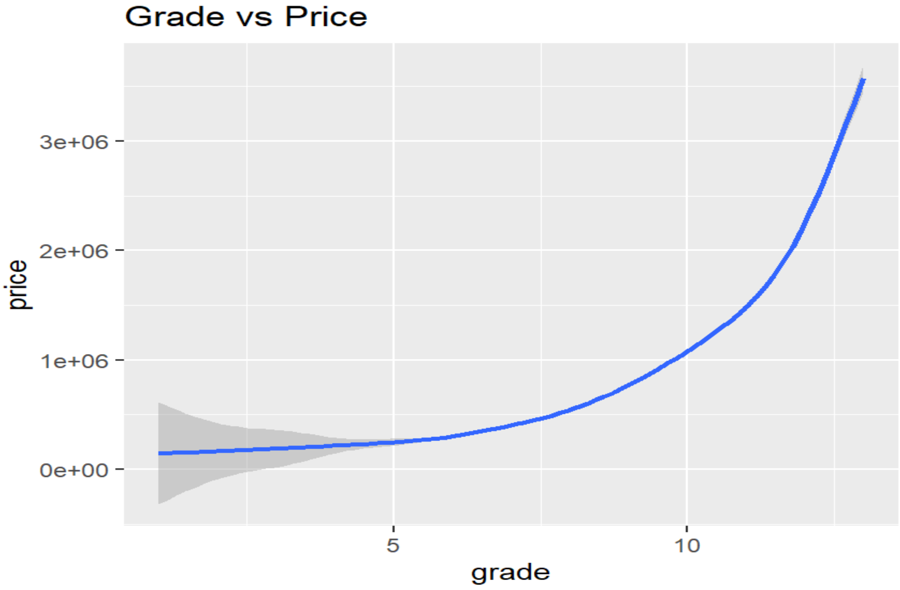

Figure 2. Correlation between Price and Grade

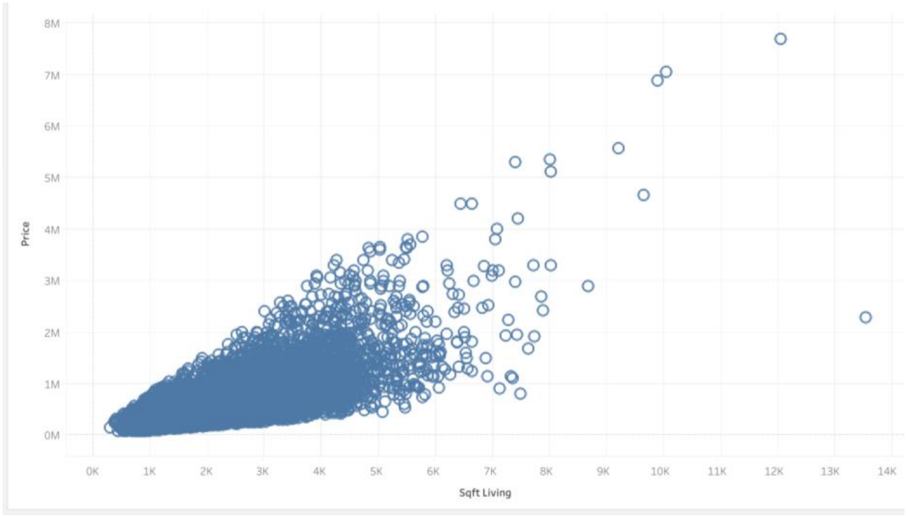

Figure 3. Correlation between Price and House size

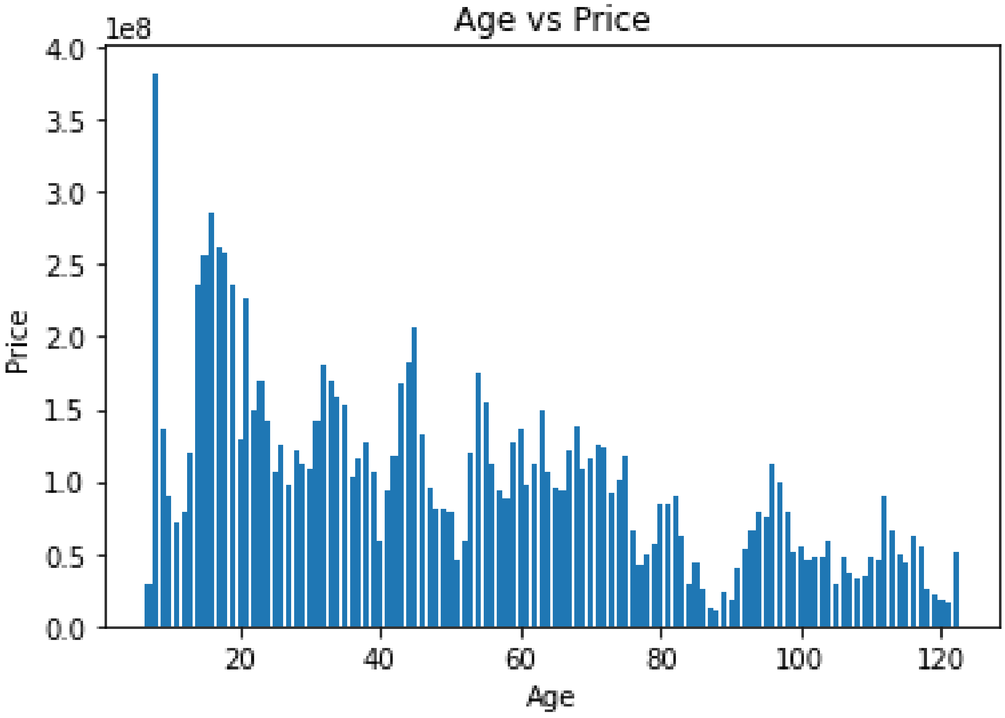

Figure 4. Correlation between Price and Age of a house

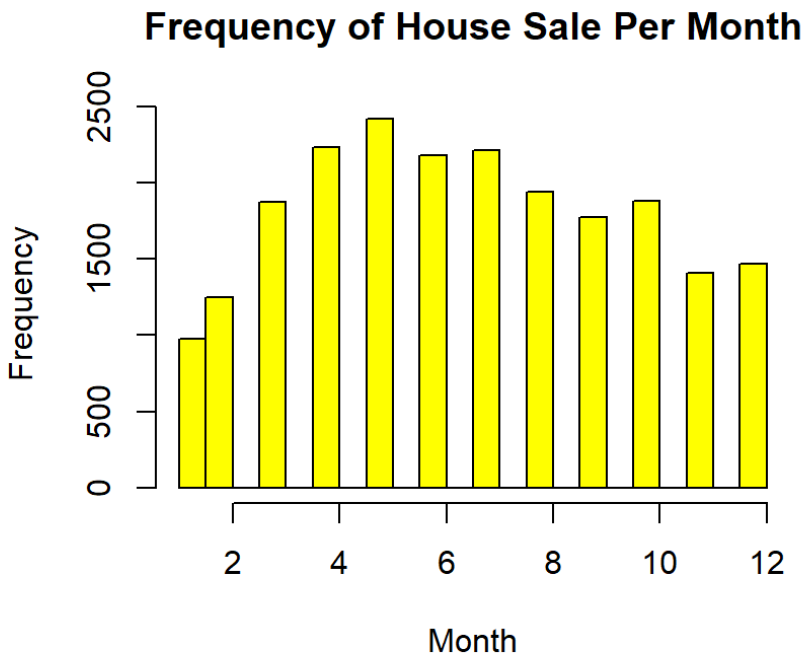

Figure 5. Frequency of House Sale per Month

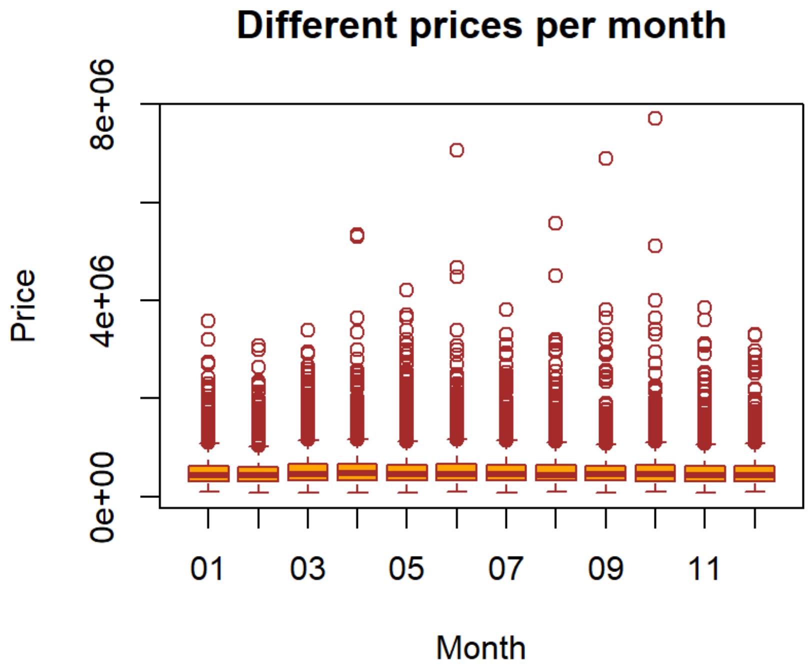

Figure 6. Different prices per month

The observation from the correlation matrix in Figure 7 show that:

- House prices have a high positive correlation with number of bathrooms, square foot living, grade, square foot above, and square foot living 15.
- Prices have low positive correlation with number of bedrooms, floors, waterfront, view, square foot basement and latitude.
- Prices have non-significant relationship with square foot lot, condition, year built, year renovated, zip code, longitude, square foot lot15, age, and renovated.
- Square foot above, square foot living 15, bathrooms, bedrooms, grades, and square foot above shows high positive relationship with square foot living and may explain the same variation in price as square foot living.

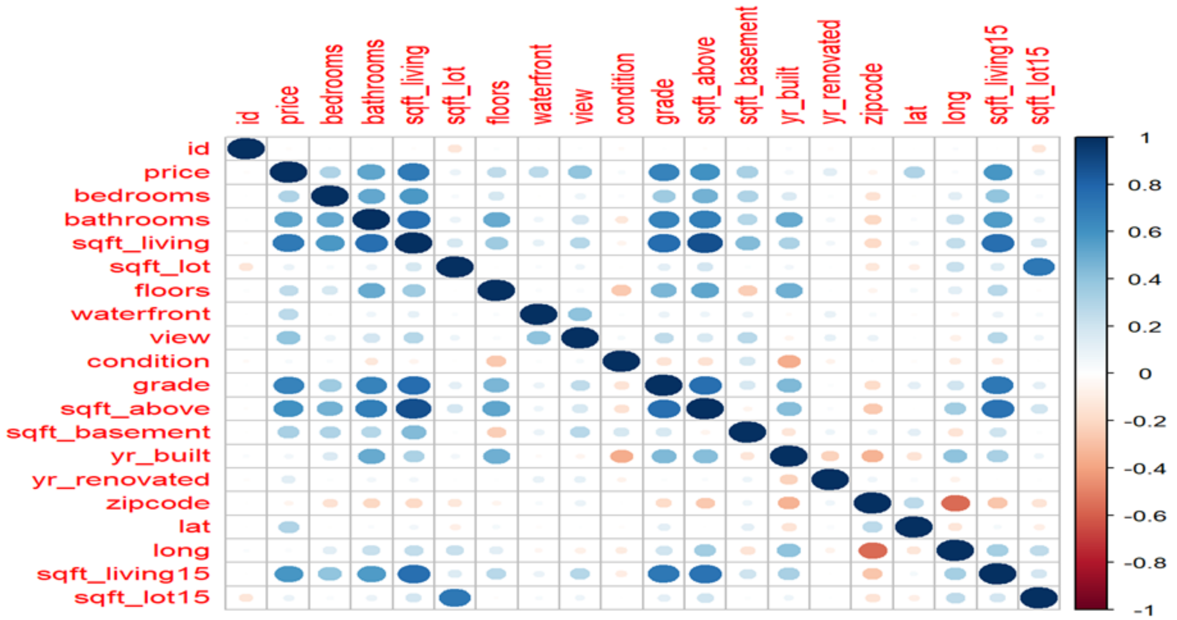

Figure 7. Correlation matrix among features

5. Study 2

5.1. Model Selection

House value of King County was analyzed and forecasted by Lasso Regression, Extreme Gradient Boost and Random Forest method. 17,289 samples were treated as the training set, 4,323 samples were treated as testing set.

5.2. Lasso Regression

Lasso is one of the variations of Linear Regression, which try to stabilize it, and make it more robust against outliers, overfitting and more. Lasso Regression allows us to eliminate some features out of our dataset and only keep the important ones, with this move, our model still perform exactly the same as before. Hence, Lasso Regression is really helpful when we want to minimize the number of features our model should use.

In this paper, I first scaled x, since Lasso performs best when all numerical features are centered around 0 and have variance in the same order. Then I used the cross-validation function RepeatedKFold, setting parameter to 10 folds, repeated 3 times with random state equals to 1. For the Lasso Regression, I used the LassoCV function in Sklearn package and set the parameter to the following:

- Alphas= arange(0, 0.01,1)
- Cv=Cv
- N_jobs=-1

The model performed not too good with an accuracy rate of 70%, MSE is around 38,735,410,323, RMSE is around 196,813, and MAE is around 125,074. Figure 8 shows the Lasso prediction in X-axis, and the actual price in Y-axis for the testing data.

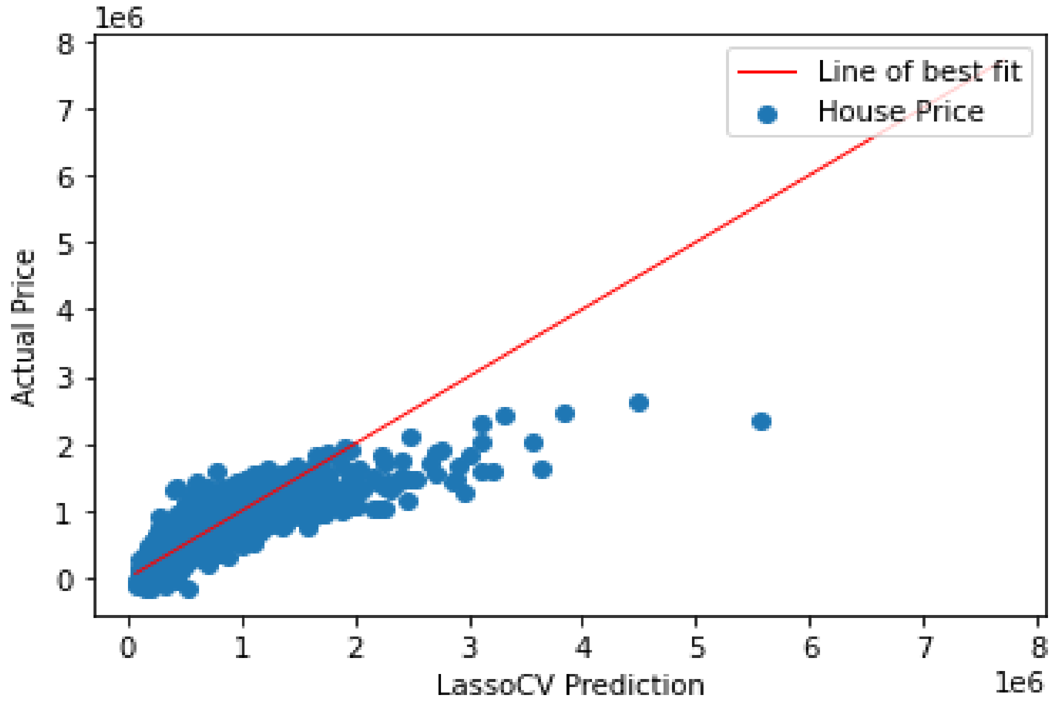

Figure 8. Lasso Regression

5.3. Extreme Gradient Boost (XGBoost)

XGBoost is a powerful approach for building supervised regression models. XGBoost expects to have base learners which are uniformly bad at the remainder so that when all the predictions are combined, bad prediction cancels out and good prediction sums up to form final good prediction.

In this paper, I utilized the XGBRegressor from xgboost open-source package. Before fitting the model with the training set, I first used the HalvingGridSearchCV function to find the best parameter for my model, then I set the parameter to the following:

- Learning_rate=0.5
- Max_depth=7
- N_estimator=12
- Cv=50

When applying the XGBoost method, the model achieved a higher accuracy rate of 88,4%, with MSE equals 14,888,762,423, RMSE equals 122,019 and MAE equals 71,109. Figure 9 illustrates the performance of this model on testing set where

X-axis is the prediction result and Y-axis is the actual housing price.

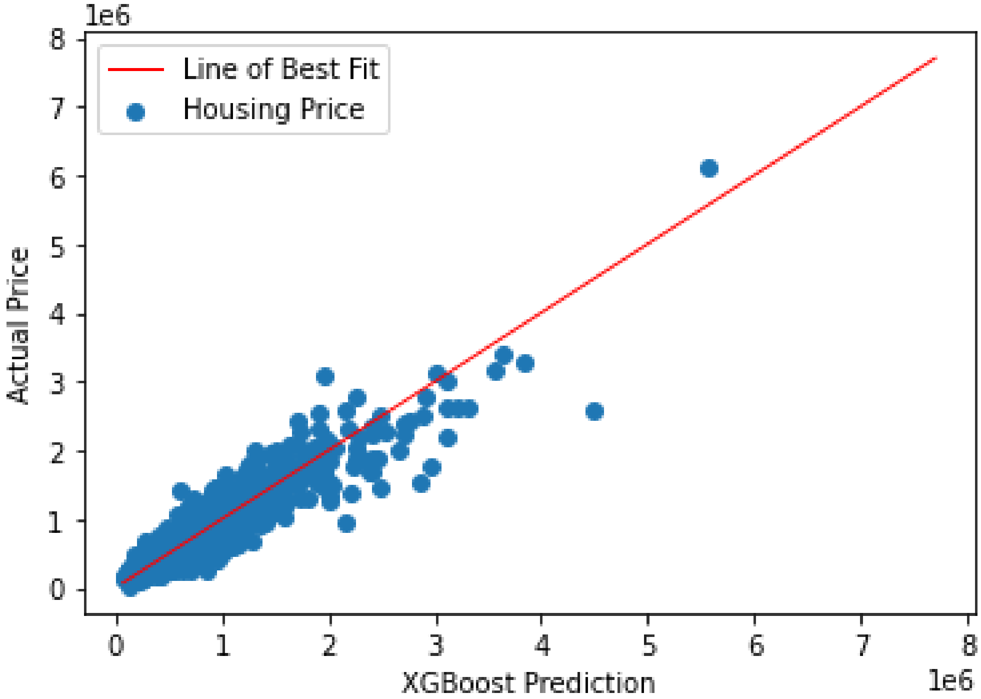

Figure 9. XGBoost Regression

5.4. Random Forest

Random Forest is an ensemble supervised learning machine method for classification and regression that combines numerous decision tree predictions to produce a more accurate final forecast. The low correlation between models is the key of Random Forest's effectiveness. The explanation for this amazing effect is that the trees protect each other from their individual errors.

In this paper, I set the parameter to the following:

- N_estimators =10
- Random_state = 1

This model returned the accuracy rate of 87.5 % with MSE equals 16,361,106,708, RMSE equals 127,910 and MAE equals 70,829. Figure 10 illustrates the performance of this model on testing set where X-axis is the prediction result and Y-axis is the actual housing price.

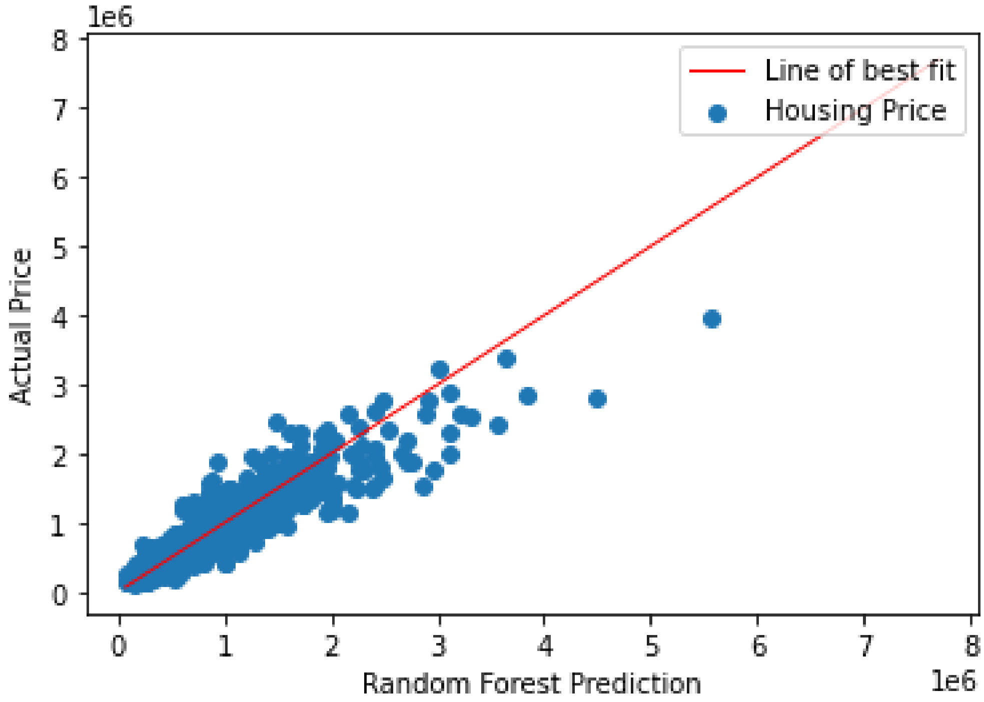

Figure 10. Random Forest

6. Model Evaluation and Selection.

Table 2 Model Evaluation Result

Model	MSE	RMSE	MAE	Accuracy
Lasso Regression	38,735,410,323	196,813	125,074	70.4%
Extreme Gradient Boost	14,888,762,423	122,019	71,109	88.5%
Random Forest	16,361,106,708	127,910	70,829	87.5%

From Table 2, we can conclude that Extreme Gradient Boost performs the best among the other two models. It has the highest accuracy rate, also the lowest MSE. Extreme Gradient Boost demonstrates a great capability of precise prediction and does not show any tendency of overfitting; hence, I choose this model as the final model used to predict housing price in King County.

7. Conclusion

Shopping for houses is an important step in many people's lives. Everyone wants to buy low and, possibly in the future, sell high. Figuring out what traits affect the price of houses in what ways can go a long way in learning how to shop for houses and what strategies to employ to maximize gain while minimizing loss. So, what have we learned?

1. Location matters! The location of the house seemed to play a role in the price greatly, with houses at waterfront and near Medina, Clyde Hill and Mercer Island being more expensive than the other.
2. Bigger and newer is better, for the most part. The larger living area, the lower the age a house, the higher housing price, also grade does matter too.
3. Prices are stable during the year, but April and May have the highest sales volume, so maybe we can create our marketing campaign in February or March in order to sell more houses.

For the predictive model, Extreme Gradient Boost outperforms Lasso Regression and Random Forest.

8. Future work

I didn’t delete any outlier in this dataset since in my perception there was some meaning behind the outlier data, but maybe delete some might contribute to a better model in the future.

References

David Berger, Veronica Guerrieri, Guido Lorenzoni, Joseph Vavra. HOUSE PRICES AND CONSUMER SPENDING. NBER WORKING PAPER SERIES, Working Paper 21667. Get Link

H. Kusan, O. Aytekin, and I. Özdemir, “The use of fuzzy logic in predicting house selling price,” *Expert Systems with Applications*, vol. 37, no. 3, pp. 1808–1813, 2010. Get Link

Wilson, I.D., S.D. Paris, J.A. Ware and D.H. Jenkins, 2002. Residential property price time series forecasting with neural networks. Knowledge-Based Syst., 15: 335-341. Get Link

Wang, J. and P. Tian, 2005. Real estate price indices forecast by using wavelet neural network. Compute. Simul., 22: 96-98. Get Link

Mundy, B. and J.A. Kilpatrick, 2000. Factors influencing CBD land prices. J. Real Estate, 25: 28-29. Get Link

Tejvan Pettinger. How the Housing Market affects the economy. Economicshelp. Get Link

Why Machine Learning Matters? Get Link

Boris Giba. Lasso Regression Explain, Step by Step. Get Link

XGBoost for Regression. GeeksforGeeks Get Link