Predicting King County House Price using Machine Learning

Author

Yu-Ting Chen
YuSha Wang
Conor Reiland

University of Washington

Contents

I Introduction

Previous Work

The main data source for this study is a house pricing dataset from Kaggle. Kaggle is a website geared towards data science and statistical research, boasting the largest community of data scientists and machine learning specialists in the world. Because of this, our dataset has been analyzed from various angles with well documented approaches and solutions. We have identified a number of unique or interesting approaches to the dataset as the basis for our continued analysis of the dataset.

- **Predicting House Prices:**

  The author predicted house prices using KNN regression, polynomial regression, and Pearson correlation. The author fit different complex models to the dataset to determine which model provides the highest predicting power. [11] We inspired from the general approach from this paper. So, first, we tried to do fundamental analysis such as correlation. Then we introduced two advanced models to optimize model performance.

- **House Sales Visualization :**

  The author is interested in data visualization and uses bar plots, correlation charts, pie charts, word clouds, box plots and maps to show the distribution of house pricing.[5] Our project also leverages data visualization to accurately describe data distributions by selecting the best presentation of information instead of presenting every possible graph.

- **Anisotropic Feature Ranking RFE, Random Forest, linear models:**

  The author provided an explanation and application of different feature ranking methods, namely Recursive Feature Elimination (RFE), Stability Selection, linear models as well as Random Forest.[3] Since our project includes machine learning, it is beneficial for us to examine the feature importance in order to have a better understanding of the models. There is more interpretation provided in our paper.

- **Multiple Linear Regression:**

  The author used backward elimination to check the usefulness of dependent variables, similar to our forward selection of certain variables to improve predicting power of the models.[9]

The previously listed related studies are only based on the data provided by Kaggle, with no additional data integration and background analysis. While these served as a good starting to begin analyzing the data, we found it necessary to look beyond research done with just our dataset. In our research into previous work, we also identified studies using different datasets that leverage unique approaches and combine several analysis techniques to produce more robust models.

- **Discovering the Hidden Structure of House Prices with a Non-Parametric Latent Manifold Model :**

  Researchers in the math and economics departments at New York University conducted a study on house pricing using a large dataset of house sales in the greater Los Angeles area.

They analyzed characteristics of houses as well as the factors that determine the desirability of a neighborhood. To do this, they combine parametric and non-parametric models, training them simultaneously using an EM algorithm. The resulting model performs better than the parametric and non-parametric models individually.[4] Similar to this study, our research approach also evaluated data outside of physical characteristics. By the result shown in the NYU paper, we are confident that additional related factors can improve predicting power.

- **The Journal of Real Estate Finance and Economics :**

  The author combines a traditional OLS regression with a spatial correlation technique called kriging to determine the price of a house based on characteristics of the house as well as prices of nearby houses. Like the previous example, this dataset also attempts to assess the impact of the environment in which the house resides. These environmental factors ended up playing an important role in our final models.[12]

Motivation

In the United States, house pricing is an important economic metric, with highs and lows following those of the national economy, occasionally influencing the national economy. For example, the 2008 recession, the biggest recession since 1945 was largely caused by a crash in the house market in which house prices fell by in case the Case-Shiller Home Price Index fell by 34%. [14] Owning a house is a major step in achieving financial independence and stability in the United States. Home ownership has long been an important part of the American Dream, so naturally, the price of houses is a concern for many Americans.

House pricing is an especially relevant topic in King County with the recent influx of tech giants building new buildings and hiring top talent with high salaries. This among other factors have contributed to a high amount of gentrification, and an increase in median house price in Seattle. Between October 2016 and June 2018, Seattle led the nation in median house price with an increase of about $200,000 in that time, topping out at $830,000.[14] With such a sharp increase in price, it is more relevant than ever to study the factors that contribute to the price of a house. There are many factors that contribute to the price of a house, some more obvious than others. Size and geographic location are obvious indicators of house price, as well as the number of bedrooms and bathrooms, the view from the house, surrounding houses and businesses, the square footage of the living room and the basement, year of renovation, etc. The goal of this study is to determine factors that contribute to the price of a house in King County, in order to predict the price of a specific house.

II Research Problem and Preparation

Questions

- **Question 1 :** What relationships exist between house characteristics and house price?
- **Question 2 :**Which features of a house are the best predictors of house price?
- **Question 3:** Which model produces the most accurate predictions for house prices?

Data Source and Examination

Data Source

The data provided by Kaggle has already been leveraged to answer questions within our area of interest, so more data has been added to the original dataset to ensure that more comprehensive factors are taken into account. With additional factors taken into consideration, we hope to build a more robust and comprehensive predictive model.

Data Examination

1. **Kaggle**[11]
   At first glance, the physical and structural components of a house seem to contribute the most to the price of the house. Kaggle dataset provides detailed information on the physical characteristics of each sold unit with the selling price. The data contains 21614 rows of data, and each row is a house sold between May 2014 and May 2015. There are 21 columns with information describing the price, area, condition, and location of the unit. There is no specific information about how the data are collected so there could be bias in the data collection process. It is also important to note the ambiguity of the grade column. We are uncertain about how the scoring system is judged. However, Kaggle is the world's largest community of data scientists and machine learners. So, we can trust their authority and credence. This dataset allows us to explore the relationship between housing price and the physical components of the house. It is an excellent dataset for evaluating simple regression models.

2. **American Community Survey (ACS)**[2]
   The US census captures data on population, demographics, and housing unit estimates. We filtered information about mean/median income, population, number of people employed, level of education in the King County region from this source. We merged these data by zip code into the Kaggle house pricing dataset. So, for each row of data, we have not only information on the architectural component, but also environmental factors that may contribute to the house's sale price. The United States Census Bureau is responsible for producing data about the American people and economy. The data selected are collected by using American Community Survey 5-Year Estimates. As the Census Bureau is a government organization, the quality and credibility of the data go without question. In general, finding the perfect place to call home depends on many factors outside of the physical characteristics of a house. We hope that the additional data will allow us to better analyze and understand the factors that influence housing price. We believe these extra factors somehow relate to customers choice of selecting a house and so further influence the demand as well as the housing price.

3. **King County**[8]
   King County.gov provided self-collected data for King County. We use the neighborhoods defined by king county to categorize relative houses location as well as their general neighborhood environment. Our neighborhood data are collected by the county demographer and other staff from King County Dept. of Transportation, Metro Transit Division, GIS Group. These reports draw on national and local sources, and they provide interpretive analysis of the application of these data sources within King County. As a government organization, the data is trustworthy and accurate. Neighborhood name has been included in the dataset as

a more granular location variable in the zip code. Based on previous research (namely the NYU study), we suspect that neighborhood will play a role in house price.

Research Approach

1. **Collecting data and data cleaning**.
2. **Feature transformation**.
3. **Analysis of variables**.
4. **Building the model and making predictions**.

Research Preparation

Data Wrangling

The Kaggle data with house prices and characteristics was already in a usable format and required no wrangling. Data from the ACS datasets populated columns 20:26 in the final dataset. First, data was read into an R script, filtering the data to columns with zip code and the total columns for relevant metrics. Columns were then renamed in order to be more readable and descriptive, eg. HC01_EST_VC13 was renamed to median_income. The first row, containing descriptions of corresponding columns was dropped in each data.frame to allow for easier wrangling. Next, columns were converted to numeric from character to facilitate merging into the original Kaggle dataset. The new data.frames containing ACS data were then joined by zip code into the Kaggle dataset.
The KC_neighborhood data came from a .shp file which contains polygons that represent boundaries for neighborhoods in king county. The rgdal R package was used to read the .shp file into a SpatialPolygonsDataframe (SPD). Using a custom function GisToDf(), our data was then copied and converted to an SPD, using the lat and long columns as the coords object in the new SPD. The function then merges two SPDs on their coords columns, so that each column in the SPD with our house price data now contains information regarding which polygon the coordinates for the house were contained in, as well as the name associated with the polygon (the neighborhood name). The data from the SPD was then merged back into the original dataset, and NA values, houses not contained in defined neighborhoods, were replaced with No Neighborhood.

Data Cleaning

In the data cleaning stage, three main problems need to be taken care of. First, the dataset contains a large number of NA values. Second, many of the columns are unrelated to our specific analysis or have only been used to join additional data. Last, outliers must be removed to normalize variables and improve the accuracy of models. Following the statistics definition, an outlier is defined as a point which falls more than one point five times the interquartile range above the third quartile or below the first quartile. In this dataset, the outliers, according to the above definition, will be removed. However, this dataset includes categorical variables, which is not suitable for removing the outliers, as they are organized by level. Consequently, outliers have not been removed from columns containing categorical data. In the data cleaning section in Appendix, Figure 9 is the boxplot that shows the distribution of each variable before and after removing the outliers. Before cleaning the dataset, the total number of rows is 21613, and after cleaning the

data, the total number of rows is 14145. The difference of the data with and without outliers can be seen in Appendix 1, Figure 9.

III Basic Statistical Analysis and Data Exploration

Summary Statistics

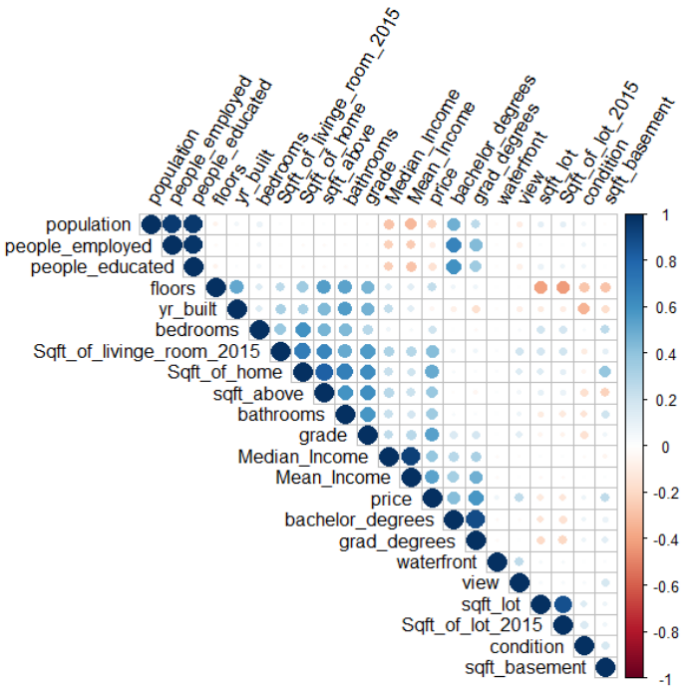
Through exploratory data analysis, we have identified key variables and summarized relevant statistics for each (See the entire data in Appendix,Table 4):

- House price in King County occupies a wide range from $78000 to $1127000 in the year of 2014-2015 with an average price of $425715.93.
- The median income on a yearly basis in King County has a range from $42965 to $120970 in the year of 2014-2015 with an average price of $76411.51.
- Comparing the income and the housing price in King County area, results show a large gap between income and house price. This indicates that house prices have become a heavy burden to the people who live in King County, increasing the probability that people will rent a house instead of purchasing one. Otherwise, there is a great chance that people will need to borrow money from the bank in order to own a house.
- When comparing the number of bachelor's degrees to graduate degrees, results indicate that on average, there are twice as many people with bachelor's degrees than graduate degrees. The following reasons may cause this: first, the living expense of the King County area may be too high for people to afford higher education such as graduate degrees or doctoral degrees. Second, people may be more likely to prioritize short-term investments over a long-term investment like a graduate degree. Last, since employment levels were high in each quantile for education levels, it is reasonable to speculate that education level is not a hindrance to finding a job in King County.
- On average, the houses in the King County area are built around 1971 - 1972, which means that a lot of the houses were more than 47 years old.
- Waterfront is a binary variable, which shows 1 when there is waterfront and 0 otherwise. On average, the house in King County does not have waterfront in the neighborhood area.
- The value of the grade is decided based on the King County's grading system, which on a scale of 1 to 13. The median value for the houses' grading is 7, which means the quality of the house in King County are averagely higher than expected since the median of the entire grading system is 6.5. A similar trend can be observed in the condition variable as well.
- For the variable "view" in the dataset, results have shown that the median value is 0 and the max value is 4, meaning most houses in King County are without a "view" of high enough quality that it may impact the price.

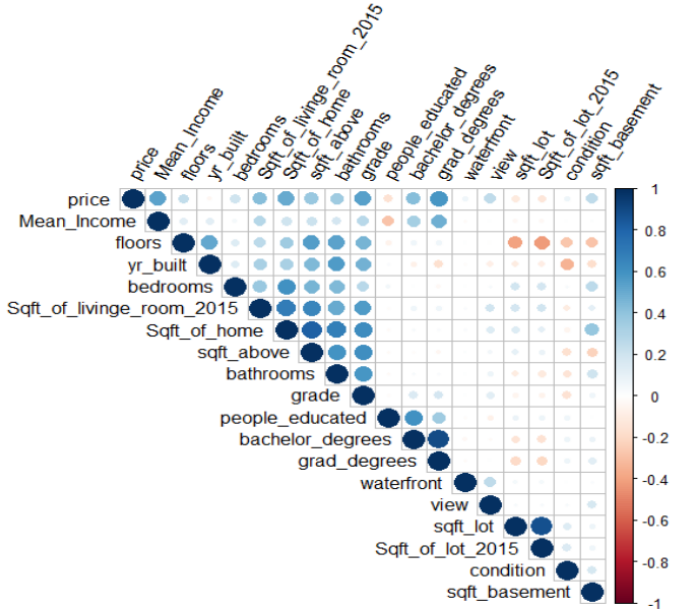
Correlation

The correlation plots describe the relationships between the explanatory variables that may contribute to houses price in King County. As shown in Figure 1 and Figure 2, a positive correlation is

shown in blue color, and the negative correlation is shown in brown color. The darker color implies a higher correlation and the lighter color implies a weaker correlation between variables.



(1) Original Correlation Plot



(2) After removing highly correlated columns

Figure 1: Correlation plots of explanatory variables

In Figure 1, several explanatory variables share a highly collinear relationship. For instance, the median income and the mean income; the number of people who earned a graduate degree and the number of people who earned a bachelor degree; the number of people who are employed

and the number of people who attend school in 2011 - 2015, etc. This dataset has run into a multicollinearity problem, which may have the following drawbacks [10]

- Many of the coefficients of the regression model may not be precisely estimated due to multicollinearity.
- The standard errors are likely to be high.
- In different samples, multicollinearity may change the sign and the magnitude of each coefficient.
- It is meaningless to assess the relative importance of the independent variables in explaining the variation caused by the dependent variable.

Although there are no two variables that result in perfect collinearity, there are still some explanatory variables which share a correlation around 0.9. In order to lower the chance of generating problems because of collinearity, some of the variables will be removed during data cleaning.

Model selection using AIC and BIC

AIC and BIC have been chosen in favor of R squared for model selection because of drawbacks with R squared with our dataset. As the number of variables increase, so too does the R squared value. Moreover, R square is a measure of explained variance in the response variable by the predictors, and it does not necessarily prove a better fitting model. On the other hand, the AIC and BIC model uses the negative log likelihood in order to penalize parameters that are related to the RSS, to improve the fit of the model. In other words, if a model fits the data better, it will have a lower value. Using the forward selection method, the upper limit of the AIC, BIC value is the set of all variables from the dataset, and the lower limit corresponds to no variables at all, and the lower limit by none of the explanatory variable is included. The difference between AIC and BIC are shown below: [1]

$$\text{Model Selection} \begin{cases} AIC = -2 \cdot ln(Likelihood) + 2k \\ BIC = -2 \cdot ln(Likelihood) + k \cdot ln(N) \end{cases}$$

Where k is the number of degree of freedoms of the model and N is the number of observations. The results from AIC and the BIC selection give the same model as shown below.

$$\begin{aligned} \text{Price}_{(House)} =& \beta_0 + \beta_1 \times \text{graduate degrees} + \beta_2 \times \text{sqft of living} + \beta_3 \times \text{people educated} + \beta_4 \times \text{view} \\ &+ \beta_5 \times \text{grade} + \beta_6 \times \text{year built} + \beta_7 \times \text{condition} + \beta_8 \times \text{bachelor degrees} + \beta_9 \times \text{bathrooms} \\ &+ \beta_{10} \times \text{sqft of living in 2015} + \beta_{11} \times \text{waterfront} + \beta_{12} \times \text{sqft above} + \beta_{13} \times \text{sqft of lot in 2015} \\ &+ \beta_{14} \times \text{floors} + \beta_{15} \times \text{bedrooms} + \beta_{16} \times \text{sqft lot} + \epsilon_i \end{aligned} \tag{1}$$

Since some of the variables in the model are categorical and others are numerical, the generalized linear regression model is then applied. The β_0 represents the intercept of the regression model, which is the value when all of the explanatory variables are zeros. If the explanatory variable is a numeric variable, the coefficients of the linear models indicate the increase or decrease in the response when increasing one unit of certain explanatory variable and holding other variables constant. These can be interpreted as the following (See the entire result in Figure 14 in Appendix):

β_1: If the number of people who earn a graduate degree increases by one, then the price will increase by β_1 dollar.

β_2: If one unit of square footage of the house increases, then the price will also increase in β_2 dollar.

β_3: In 2011-2015, if the number of people who have been educated increases by one, then the price will increase by β_3 dollar.

β_8: If the number of people who earn a bachelor degree increases by one, then the price will increase by β_8 dollar.

β_9: If the number of bathrooms increases by one, then the then the price will increase by β_9 dollar.

β_{10}: In 2015, if the living room area increases by one unit, then the then the price will increase by β_{10} dollar.

β_{12}: If the square footage of the house apart from the basement increases by one unit, then the then the price will increase by β_{12} dollar.

β_{13}: In 2015, if the Lot size area increases by one unit, then the then the price will increase by β_{13} dollar.

β_{14}: If the total floors in the house increases by one unit, then the then the price will increase in β_{14}

β_{15}: If the number of bedrooms in a house increases by 1, then the then the price will increase in β_{15} dollar.

β_{16}: If the square footage of the lot increases by 1 unit, then the then the price will increase in β_{16} dollar.

For categorical variables such as grade, condition, view, waterfront and year built, their coefficients explain the relationship with the reference level. In the full model, for instance, the coefficients of the variable "view" are than shown as $\beta_{4_1}, \beta_{4_2}, \beta_{4_3}, \beta_{4_4}$ instead of simply β_4. This is the same for other categorical variables. In the following interpretations, only the first sub-coefficient will be explained; other sub-coefficients will have a similar style of interpretation. The full table can be seen in Appendix, Figure 14.

β_{4_1}: The base level is view = 0. On average, we expect the price of the house with a view rated at one to cost β_{4_1} more compared to that of a house with a view rated zero.

β_{5_1}: The base level is grade = 4. On average, we expect the overall grade given to the housing unit with a grade equal to seven to cost β_{5_1} more than the house with a grade level equal to six.

β_{6_1}: The base level is year built = 1900. On average, we expect a house built in 1901 will cost β_{6_1} more compared to a house built in 1990.

β_{7_1}: The base level is condition = 1. On average, we expect a house with condition level of three will cost β_{7_1} more compared to a house with a condition level of two.

β_{11_1}: The base level is waterfront = 0. On average, we expect a house with a waterfront view will cost β_{11_1} more compared to a house without a waterfront view.

Different from machine learning, the AIC model and BIC model select the best model based on the entire data instead of the training set. Therefore, we would like to build our models that are capable of testing predicting accuracy in the next section by tuning the parameters in Random Forest and XGBoost. Different from AIC and BIC, only the training set and cross-validation set will be used in constructing the models.

IV Machine Learning Predictive Modeling

Random Forest

Brief Overview of Random Forest

Random Forest is a supervised learning algorithm that is widely used in machine learning. The random forest model is very good at handling numerical features or certain features with few columns. Unlike linear regression models, random forests can capture the non-linear relationship between the explanatory and the predictor. Also, collinearity will not affect the predictive power. This algorithm is also an excellent choice for building a robust predictive model in a short time frame.

In the housing price dataset, it has numerical features such as the number of bedrooms as well as categorical features such as a grade. There is a non-linear relationship between the explanatory and price as well as collinearity between the mean and median income. Since the time for doing machine learning is limited and the dataset fit all the characteristics mentioned above. Random forest stands out as a beautiful housing price predicting model.

Random forest builds multiple decision trees and merges them to get the final predicting model. More formally we can write this models as: $g(x) = f_0(x) + f_1(x) + f_2(x) + ...$ where the final model, random forest regression g, is the sum of simple base models, decision tree, f_i.[15]

Each decision tree goes from observations about an item to conclusions about the item's target value. In one decision tree, each internal node is a test on a single attribute, each branch represents the outcomes of the test, and each leaf holds the result of the target value. For each independent variable, the data is split at several points. We fit a regression model at each split region and the model with the minimum sum of square error the specific decision tree.

Random Forest randomly selects observations to build several decision trees and then averages the results. The algorithm searches for the best feature among a random subset of features. Therefore, the random forest can prevent over-fitting most of the time.

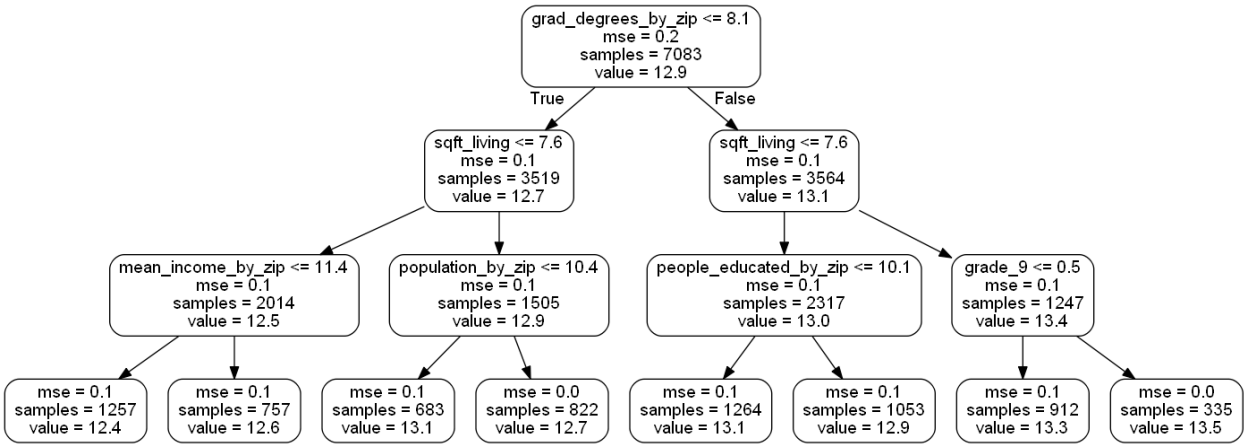


Figure 2: Decision tree snapshot

In Figure 2, the root is the first box. "The number of graduate degrees" is the variable and value on which the node is split. "Mse" is the mean square error. "Samples" is the number of samples in this node, and "value" is the predicted price (after normalization) value based on the samples inside the node. Based on this small tree, we can predict any new data point by verifying the condition in each node, choosing the branches, and then entering another node. The process repeats until we reach a terminal leaf, in this case, entering the fourth level of the tree.

Hyper-parameters

By default setting of random forest regression, the root mean square error is 0.167, which is the standard deviation of the unexplained variance between exact housing price and predicted housing price. So, choosing appropriate hyperparameters plays a crucial role in the success of the predicting model. We want to further reduce the RMSE value by changing hyperparameters since lower RMSE values indicate a better fit. However, hyperparameter tuning relies more on experimental results than theory. In this model, we used the standard procedure for hyperparameter optimization through cross-validation. Using Scikit-Learns RandomizedSearchCV method implemented in python, we select a grid of hyperparameter ranges, and randomly sample from the grid, performing 3-Fold cross-validation with 100 different combinations of values. Below is a list of hyperparameters mentioned in RandomForestRegressor. All of the values need to be decided:

- n_estimators = number of trees in the forest
- max_features = max number of features considered for splitting a node
- max_depth = max number of levels in each decision tree
- min_samples_split = min number of data points placed in a node before the node is split
- min_samples_leaf = min number of data points allowed in a leaf node
- bootstrap = method for sampling data points (with or without replacement)

By executing RandomizedSearchCV, hyperparameters are optimized by a cross-validated search over parameter settings provided below. These ranges are widely used in industry. The result will show the best set of hyperparameters that produces the model with optimum performance. It is worth noting that not all parameter values are tested, but a fixed number of parameter settings are sampled from the distributions.

After narrowing down the range for each hyperparameter by using random search, we further specify every combination of settings. The method is called GridSearchCV. It evaluates all combinations in the provided grid using a 3- Fold cross. The results are shown in Table 1:

	X	Range	Best.Choice
1	N Estimators	200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000	1200
2	Max Features	auto, sqrt	sqrt
3	Max Depth	10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, None	80
4	MinSamplesSplit	2, 4, 5, 6, 7, 10	7
5	MinSamplesLeaf	1, 2, 4	1
6	Bootstrap	TRUE, FALSE	FALSE

Table 1: Hyper-Parameter Choice and Results

These hyperparameters can predict the housing price with lowest RMSE based on our training and testing sets.[13][6]

ML decisions

During data preparation, categorical variables are converted to a numerical representation without an arbitrary ordering. This process is known as the one-hot encoding of the data. Therefore, the algorithm will not place more importance on some categories with a higher numerical value. Besides, since we are not sure of the grading and condition ranking system inside the data, these categorical columns are changed into columns without an arbitrary ordering, assuming they have equal importance.

Before making and evaluating predictions, we establish a baseline, a sensible measure that we hope to beat with our model. If our model cannot improve upon the baseline, then it will be deemed unsuccessful and discarded in favor of another model. The baseline prediction for our case is the RMSE between the mean of housing price in our test set and the real testing housing price. The baseline error is 0.415 is indeed larger than the RMSE error generated by fitting random forest. So, in this case, we are confident that the random forest regression model can improve our predicting model and results.

Results and Feature of Importance

The RMSE for predicting housing price is 0.1625 by using the best parameters from fitting the random search. The RMSE reduced by 2.69%. By re-checking the RMSE with grid search, it is 0.1622, further reduced by 0.18%. Since the amount of reduction is small, there is no need to narrow down the searching range further. Therefore, our last random forest predicting model has an RMSE of 0.162 between the predicted housing price and the actual housing price in the test set. In order to quantify the usefulness of all the variables in the entire random forest, we look at the relative importance of each feature. Essential features are more closely related to the dependent

variable and contribute more to variation in the dependent variable. These numbers offer us insight into the relative predicting power between variables.

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples.

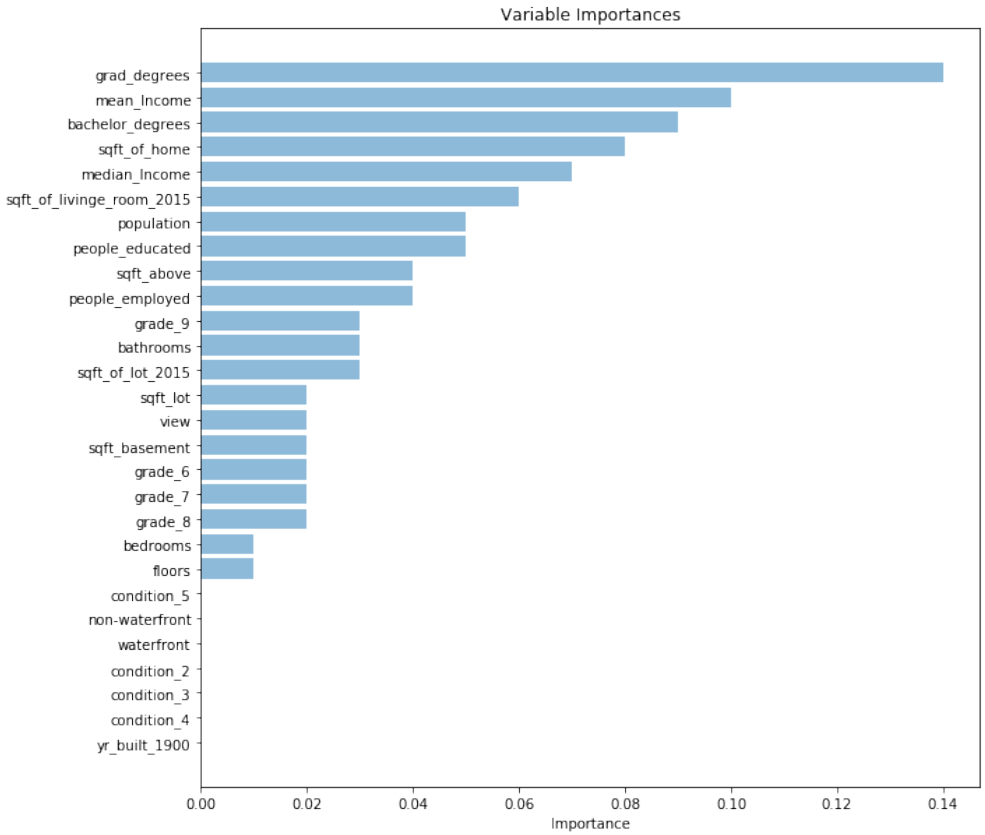


Figure 3: Variable importance (Built year after 1900 have 0 importance and are not showed on the graph)

In figure 3, Number of graduates increases prediction RMSE most if it gets deleted. So, node splits based on this feature on average result in a considerable decrease of node impurity. Thus, it is the most critical feature. From figure 3, number of graduate degrees, mean income, number of bachelor degrees, square footage of the home, and median income are the five top most important features contributing to the housing predicting model for each region. On the contrary, built year, whether facing waterfront and condition grade is not important at all. There is little relative importance for the square footage of the lot, square footage of the basement, King county grade level six to eight, number of bedrooms, and number of floors.

The link between physical characteristics of a house and its resulting sale price seems plausible. However, by comparing these variables' importance, the general conclusion is that the location and environment where the house is built are more important than the architectural factors the number of graduates and undergraduates, number of educated people, population size, and number of employed population influence the environment in which a house is built, making it more or less desirable. These variables are correlated with mean/median income as observed in the correlation

plot. So, income is also an important feature when evaluating housing price. On the one hand, it indicates that these people have the potential to earn more money to buy a more desirable house. On the other hand, it is highly likely that the place where these people live represent a higher quality of life. Location continues to be as relevant because the location is ultimately what brings value to a piece of real estate. For example, a house near the university would have a comparatively higher value than in the rural area of King County. Housing price is not only an aggregation of the physical components but of the market demand for that particular structure in that particular place.

From the machine learning process, we also understand that some physical characters of the house contribute to the house price such as the square footage of the home and whether the house is renovated. The grade applied to a house in the King County dataset as well as the condition of the house seem inconsequential in a buyer's choice to purchase a house. Lower grades do not necessarily mean a lower house price.

eXtreme Gradient Boosting (XGBoost)

Brief Overview of XGBoost

XGBoost is a algorithms that is constructed based on modification of the traditional GBDT algorithm (gradient boosting decision tree algorithm). In other words, it is also a tree-based model. Boosting is an ensemble technique where new models are added to correct the errors made by existing models. Models are added sequentially until no improvements can be made in the future steps [17].

The reason why XGBoost algorithm can surpass other algorithms outperforms to its novel scalability, and the factors contributing to the scalability are as follows:

- A tree algorithm that can handle the sparse data.
- Parallel, distributed computation speeds up the model exploration, which allows the model to learn faster through the training dataset.
- It allows us to exploit out-of-core computation, meaning that it can process data that is too large to fit into a computer's main memory at one time.

In consequence, the algorithm will not only produce results more efficiently than other algorithms with its powerful scalability but also allow the user to tune the hyperparameters in order to customize their objectives and evaluation based on their needs [17]. Unsurprisingly, this algorithm has become one of the machine learning algorithms chosen most often in winning entries in Kaggle competitions. On account of its accuracy and efficiency, XGBoost has been chosen as the main algorithm for predicting house prices in this study.

Log Transformation

Before applying this algorithm, it is necessary to check for the normality of the numeric variables transform data with a high amount of skew (absolute skewness more than 0.85) in order to hone the precision of the model. First, we compute the Fisher-Pearson coefficient of skewness:

$$G = \frac{\sqrt{N(N-1)}\sum_{i=1}^{N}(Y_i - \bar{Y})^3}{N(N-2) \times s^3}$$

G: Skewness

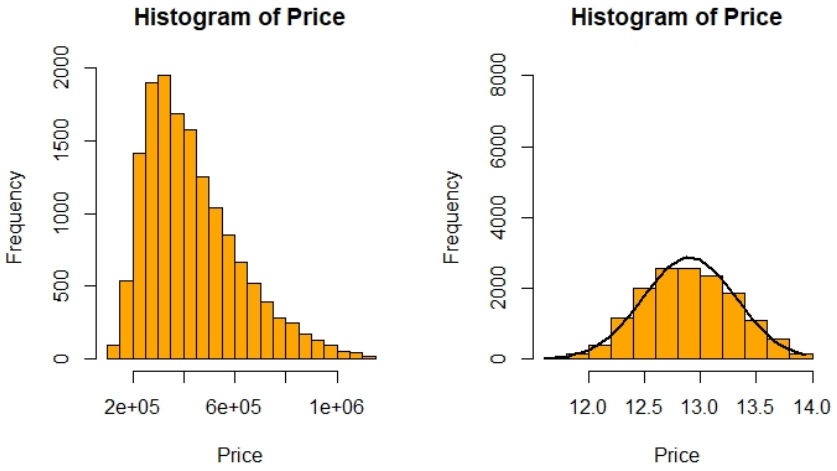
N: Size of the sample size

s: Standard deviation

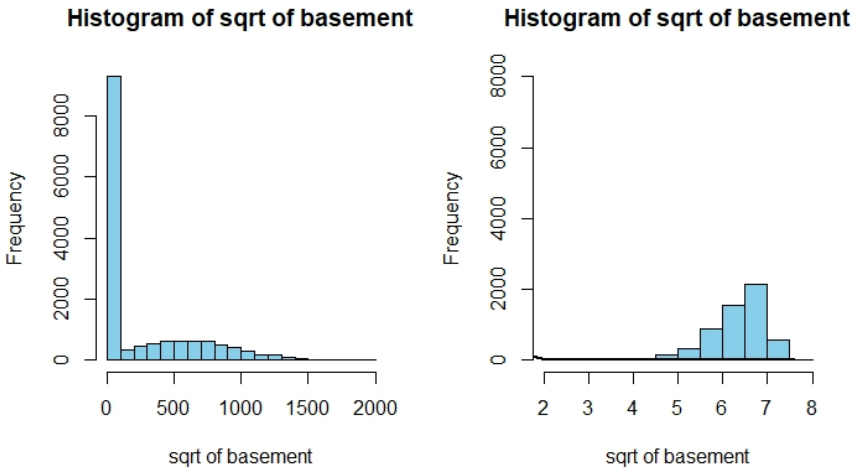
$\boldsymbol{Y_i}$**:** Sample data

$\bar{\boldsymbol{Y}}$**:** Sample mean

In the dataset, the variables sqft of basement, price have a skewness more significant than 0.85. In order to reduce the effect of valid extreme values, the log transformation is applied, since the log value will have a smaller range and less chance to excessively influencing the model. As a result, this can be valuable both for making patterns in the data more interpretable and for helping to meet the assumptions of inferential statistics.



(1) Price



(2) Sqrt of Basement

Figure 4: Histogram before and after log transformation

Hyper-parameters

After transforming the dataset, the data is then split into a training set, and testing set randomly by a ratio of 3:1, which is the same dataset used in the random forest algorithms. There are almost 20 hyperparameters in the XGBoost algorithm, and some of them are selected by interest to tune with the 5-fold Grid Search cross-validation methods:

- **learning rate:** This is also knowns as eta, which is the step size shrinkage used in an update to prevent overfitting. A higher learning rate increase computation speed, but make it harder to reach the optimum model (optimum = perfect bias/variance tradeoff). The best choice is 0.03, which means if 1 step is made at this learning rate, the step weight will be 0.03 [16].
- **max depth:** The maximum depth of a tree. As the value of this parameter gets larger, it is more likely to be overfit[16].
- **min child weight:** This corresponds to the minimum number of instances needed to be in each node[16].
- **subsample:** This is the subsampling ratio of the training instances, which will happen in every boosting iteration.
- **colsample by tree:** This is the subsampling ratio of columns when constructing each tree[16].

	X	Range	Best.Choice
1	Learning Rate	0.01,0.03,0.05	0.03
2	Max Depth	4,5,6,7,8	5.00
3	Min Child Weight	1.5,3,4	3.00
4	SubSample	0.2-1	0.90
5	colSample by Tree	0.2-0.7	0.20

Table 2: Hyper-Parameter Choice and Results

Results and Feature of Importance

With this model, the root-mean-square error is used as the criterion of the precision:

$$rmse = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{m}}$$

, which m is the sample size.

Results has shown that the rmse for the training set is 0.01154. Using the prediction model to test the test set, the resulting rmse we get is around 0.01214. From the relatively small rmse values and the proximity of the values, we can conclude the following:

- The model is not subject to overfitting.
- The prediction is accurate enough.

Feature importance has been made based on the training set as shown in Figure 5, Figure 6 and Figure 7.

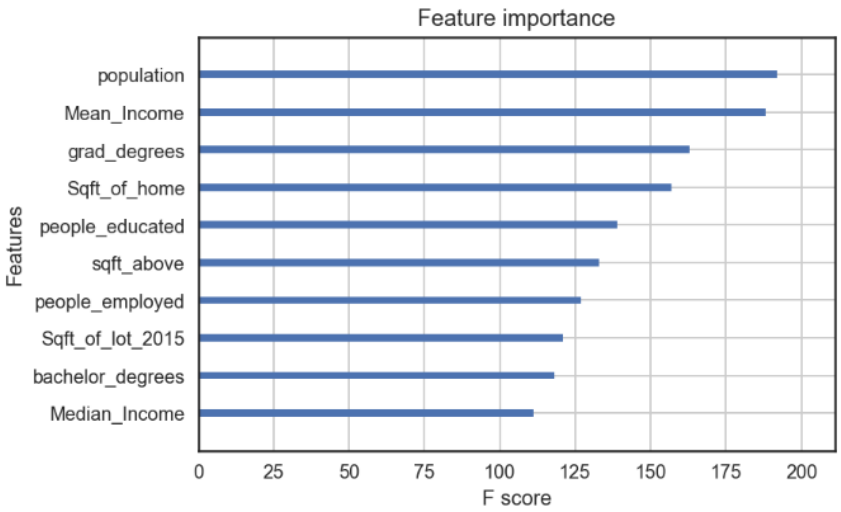


Figure 5: Feature of importance by XGBoost with importance type "Weight"

The feature of importance is calculated by the increase of the models prediction error after permuting the feature. A feature is determined to be important if permuting its values will increase the error of the model. This has indicated that the model relied on this feature for the prediction. On the other hand, the feature is regarded as not crucial if permuting its values keeps the model error unchanged. This indicates that the model ignored this feature for the prediction. As the value of specific feature shown in the plot gets higher, the particular feature is more important compared to others.

By choosing the importance type "Weight", the feature of importance will be decided by the number of times a feature is used to split the data across all trees. The top 10 features that are most important evaluate by the XGBoost are listed in Figure 5. These variables are the population, mean income ,the number of people who owned a graduate degree, the footage of the home,the number of people who was educated, square footage of house apart from the basement,number of people who were employed,the lot size area in 2015, the number of people who earned a bachelor degrees, and the median income. With this sequence, the following conclusion can be made.

- Compare to the education level in a certain area in King County, the house price is more likely to be affected by the income level in the area.
- Different from traditional concept, the square footage of the home is not as influential as the number of people who have graduate degrees.
- Consistent with the phenomenon in most big cities, as the population gets larger, it is more likely to results in a higher house price. Since King County is located in the big city, the house prices are largely affected by the size of the population.

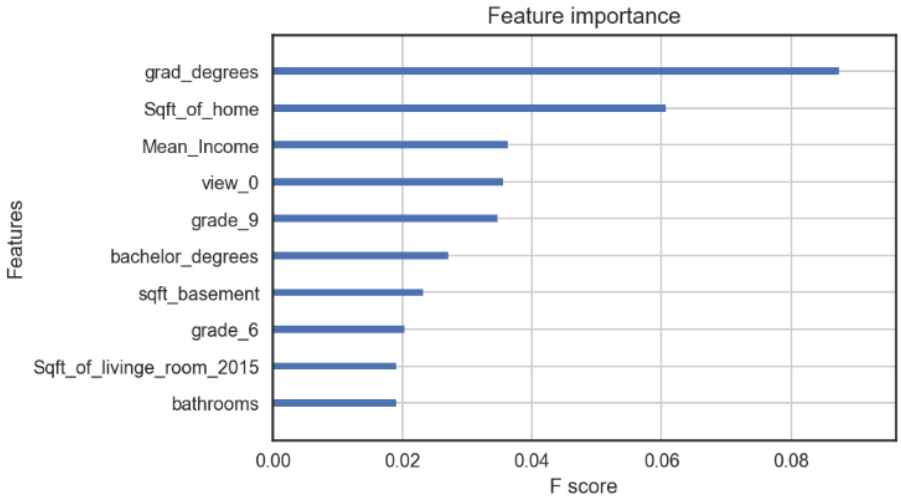


Figure 6: Importance Type with Gain

Analysis and comparison for the top 10 features with the importance type "Gain":

- The number of graduate degrees contributes to the price more than the income level.
- Compared to importance type "Weight", the population is not even listed in the top 10 factors.
- Contrary to popular belief, the size of the basement is more important than the number of bathrooms. This is an interesting finding since people more often care about individual bathrooms more than the size of the basement. One of the possible reason is that people can endure sharing a bathroom with others, which means that as long as people have access to a bathroom, they care about the size of the basement more.
- Different from Figure 5, the feature of importance in Figure 6 includes categorical variables.

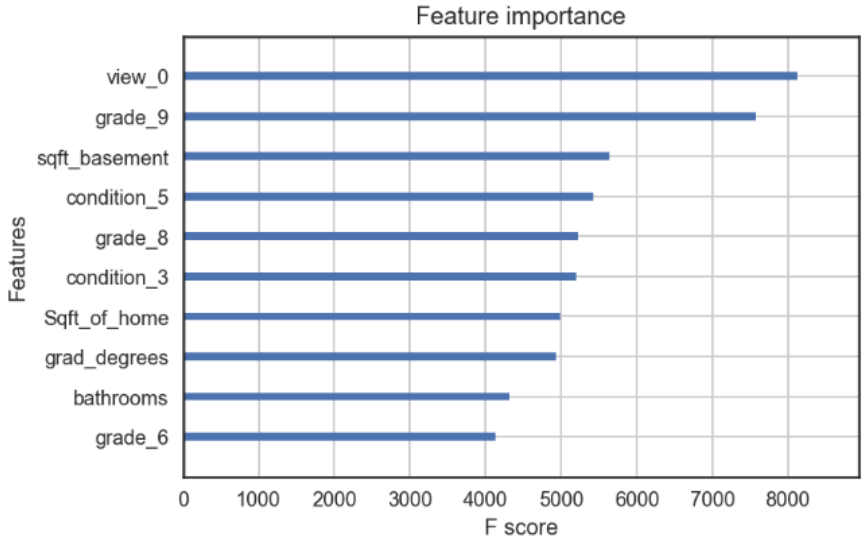


Figure 7: Importance Type with Cover

Analysis and comparison for the top 10 features with the importance type "Gain":

- Compared with Figure 5 and Figure 6, Figure 7 includes more categorical variables.
- Same as in Figure 6, the size of the basement is much more important than the number of bathroom.
- Compare to Figure 5 and Figure 6, the size of the house has lower importance in Figure 7.

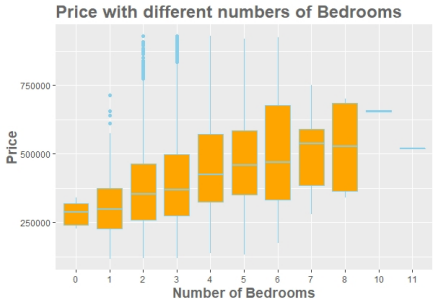
Other general importance types are "Gain" and "Cover". The type "Gain" evaluates the importance of certain factor by the average training loss reduction gained when using a feature for splitting. On the other hand, the type "Cover" measures the importance by the number of times a feature is used to split the data across all trees weighted by the number of training data points that go through those splits, which is more similar to the type "Weight"[7]. However, if the importance type is changed to "Gain" or "Cover", the top 10 important factors will change. In Figure 6, when the importance type is chosen as "Gain", the factor of tenth highest importance is the number of bathrooms, which is not listed in the importance plot that has type"Weight". This is the same for other variables, such as grade and view. In Figure 7, when selecting the importance type as "Cover", the results are much more different compared to the other two types.

Since the relative importance in the three types has enormous differences, we can conclude that there is a lack of consistency across types. In other words, it is unreliable to trust either of the important feature plots. Because having a higher assigned attribution does not mean the model relies more on that feature, and more tests should be applied.

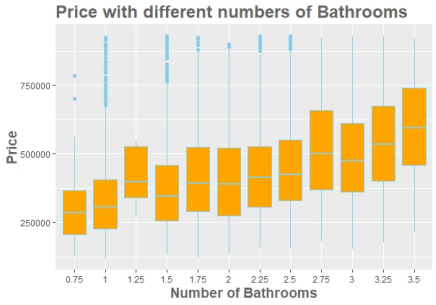
Several reasons may cause the differences in Figure 5, Figure 6 and Figure 7. The first aspect is due to the different algorithms behind the features of importance. Because three types of importance are calculated in a different way, there is inherently a possibility of producing a different, inconsistent order across algorithms. However, on the other hand, if the consistency of importance exists in such dataset, but results in a different order in three types of importance, there is also some speculations that may explain this situation. To begin with, the shuffling operation for the permutation adds randomness to the results, which may cause different results with repeated permutations[7]. Next, the collinearity properties between the explanatory variables may also be taken into account. For the XGBoost model applied in this section, the explanatory variables which share high correlations are included in the model. When two variables are highly correlated, the measure of importance may be biased. When a permutation is applied, the model may create a combination of two highly correlated variables that is unlikely to happen in the real world. Moreover, adding a correlated feature can decrease the importance of the associated feature, by splitting up the importance of both features[7]. This can be proved by Figure 8. Figure 8 is a series of boxplots that factorized the x-factor to show if different levels of certain explanatory variables can influence the price. If the boxplot in each level does not share much overlap with adjacent levels, this indicates that different levels of the explanatory variable have influenced the price. Otherwise, it cannot be concluded that the explanatory variable can impose effects on price. In Figure 8, the number of bathrooms, the number of bedrooms, total number of floor, different levels of condition, total count of viewed and built year all have boxplots that largely overlap the adjacent boxes. This indicates that they should not have a significant effect on the price. However, referring back to Figure 7, the number of the bathroom and the grade seems more important than the population, mean incomes and other variables that do not have serious overlap in the boxplots.

On the other hand, the feature of importance is different for Random Forest Algorithms and the XGBoost as well. This may be explained by the first aspect proposed earlier, which is due to the different algorithms. For the random forest, each feature is chosen at random to use for each tree,

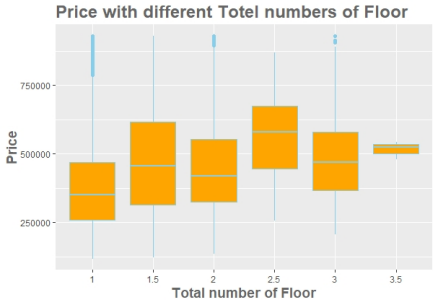
which means that that two explanatory variables that are highly correlated may be chosen to stay on the same tree. In other words, it is also possible to find a great model blindly. However, for XGBoost, the selected features are not as random as the Random Forest. It relies on tuning the hyper-parameters to find the best fit model that is not overfitted. Moreover, it extends boosting by imposing regression penalties. In consequence, the difference in features of importance in these algorithms is more likely caused by their different mechanisms.



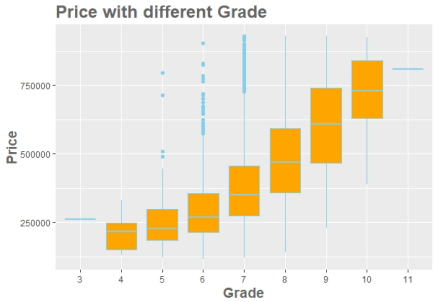
(1) Price with different numbers of Bedrooms



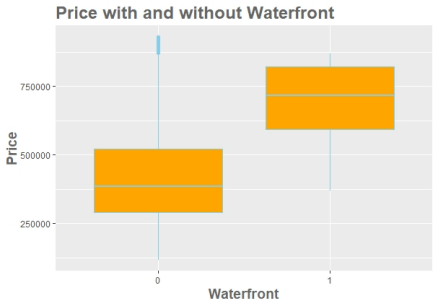
(2) Price with different numbers of Bathrooms



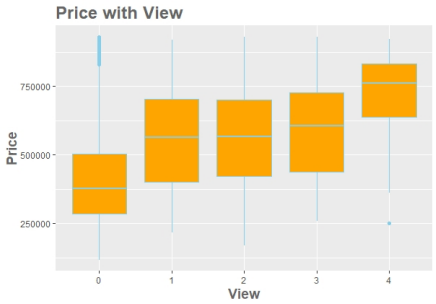
(3) Price with different Totel numbers of Floor



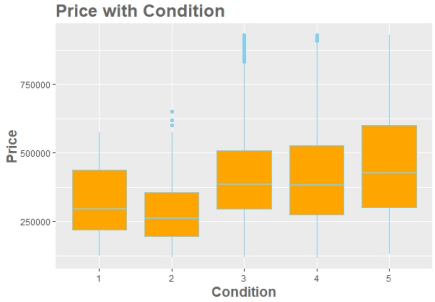
(4) Price with different Grade level



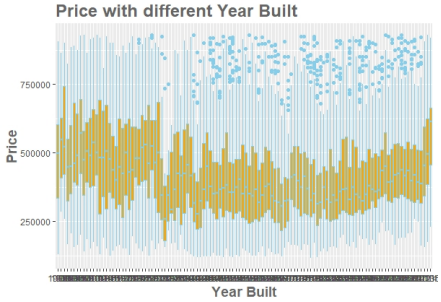
(5) Price with or without Waterfront



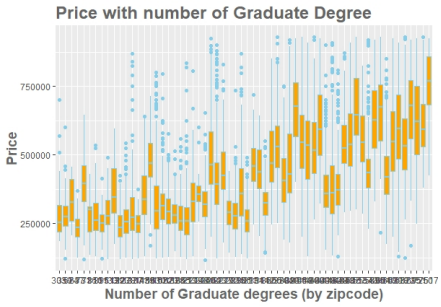
(6) Price with Total count of Viewed



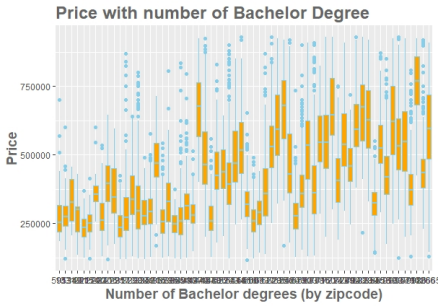
(7) Price with different Condition



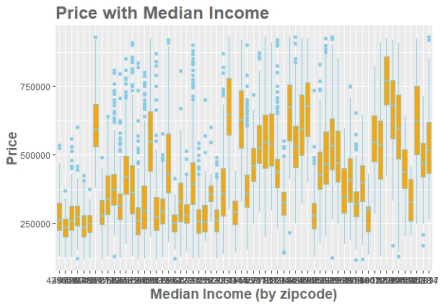
(8) Price with different Year Built



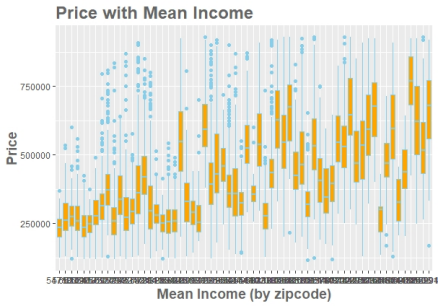
(9) Price with different numbers of Graduate Degree



(10) Price with different numbers of Bachelor Degree

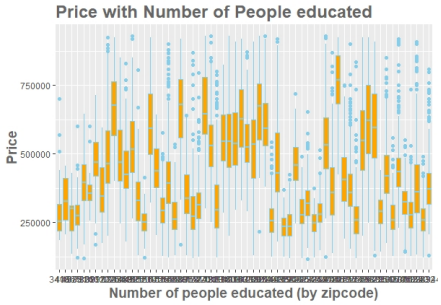


(11) Price with different levels of Median Income



(12) Price with different levels of Mean Income

(13) Price with different Size of Population



(14) Price with different numbers of people being Educated

(15) Price with different numbers of people being Employed

Figure 8: Boxplot for different levels of Factors v.s. Price

V Limitations and Discussions

Data Collection and the Nature of the Data

1. As with any predictive model, our models are unable to account for the nearly infinite confounding variables that may contribute to house prices. For instance, economic factors like changes in taxes, government policies, stock market prices, property value, inflation. Each of these can fluctuate on a monthly, weekly, or daily basis, and may contribute to minute differences in house prices.

2. We cannot account for differences in parties involved in a sale. A house sale comes down to an agreement between a buyer and seller upon a price that satisfies both parties. Each has different wants and needs when it comes to big financial decisions. A buyer may be willing to buy above market price to secure a house against many competing buyers. On the reverse side, a seller may sell below market price if he or she is in need of quick cash and unable to find a buyer immediately. Additionally, the negotiating skills of each party and real estate agents can result in differences in sale price on similar houses.

3. The dataset lacks information on the data collecting process and method. The dataset requires further and more in-depth examination. Besides, some columns have ambiguous meaning and no extra information is provided.

4. The models in this project are limited in prediction power by the scope of the dataset. The predictions produced by the models in this study only apply directly to the King County house market. Markets in different parts of the world, country, or even state could be vastly different based on many complex factors.

5. The study is limited by time and budget. With only a limited time for data collection, cleaning, and statistical analysis, it is difficult to cover all bases and deal with potential confounding variables. Specifically, difficulties were combining new datasets as much of the data was uniquely identified by the houses address which was missing from our dataset. In order to reverse geo-locate address from.

Statistical analysis and Machine Learning

1. The high collinearity in the dataset may influence the accuracy of the AIC and BIC when selecting the best fit model. Even though the columns that correlate 0.9 are removed, there are still some high correlated relationships, such as columns that share a correlation around 0.87.

2. Using the simple statistics definition to remove the outliers may not be accurate enough, other methods such as bagplot should be applied next time, since it may be more accurate to determine the outliers based on the plot instead of just from the mathematics formula.

3. For correcting the skewness, log transformation will not always have an ideal result for normalizing the data, a more advanced method such as box-cox transformation may be more accurate.

4. Random forest despite its fast processing speed can produce varied results based on the randomly selected trees. The result can be even less accurate by using random search cross validation since only a subset of hyper-parameters is selected.

5. It took a large amount of time when tuning the hyper-parameters of the XPBoost. So, a GPU accelerator may be needed. With a larger dataset, running the algorithm on a server may be necessary.

VI Plan

Work Distribution

- **Conor Reiland :**
    1. Previous Work
    2. Research Questions
    3. Motivation
    4. Data Wrangling
    5. Limitation and Discussion
- **YuSha Wang:**
    1. Data Source and Examination
    2. Research Approach
    3. Random Forest Algorithm
    4. Limitation and Discussion
    5. Appendix Creation
- **Yu-Ting Chen:**
    1. Data Cleaning
    2. Summary Statistics
    3. Correlation
    4. AIC and BIC model selection
    5. General Linear Regression Model
    6. XGBoost Algorithm
    7. Limitation and Discussion
    8. Future Plan
    9. Appendix Creation
    10. All the format of Latex.

Future Plan

In future iterations of this study, many improvements can be made. As far as data collection, the model should be trained based on more data instead of just the data from King County. Next, the data should also be from a broader range of years, since the dataset only encompasses sales between 2011 and 2015. Moreover, more related factors that may affect the house prices should be investigated in order to increase the precision of the predictions. For instance, text data such as the news should also be included, since some of the other effects like financial crises may also affect the house prices in certain areas. Additionally, more models can be applied to the prediction. For example, principal component analysis can be included to lower the influence of multicollinearity among explanatory variables. Fourth, without time limitation, more hyper-parameters can be tuned in order to achieve lower rmse without over-fitting. Last but not least, with more data, the model should be built and trained on a server instead of a PC. This is not only of the burden imposed on them but also running on the server can save time and allow for more analysis iterations.

VII Appendix

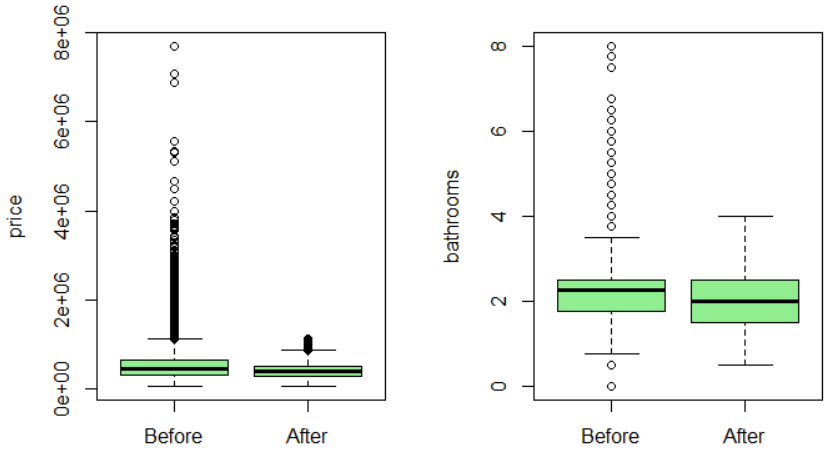
Data Column Definition

Column Name	Column Meaning	Data Type	Measurement Scale
Price	selling price for the house in dollars	integer	ratio measurement
Bedrooms	Number of bedrooms per house	integer	ratio measurement
Bathrooms	Number of bathrooms per bedroom	integer	ratio measurement
Sqft_living	square footage of the home	integer	ratio measurement
Sqft_lot	square footage of the lot	integer	ratio measurement
Floors	Total floors in a house	double	ratio measurement
Waterfront	House which has a view to a waterfront denotes as (1) if not denotes as (0)	categorical	nominal measurement
View	Has been viewed	integer	ratio measurement
Condition	How good the condition is (1-5)	categorical	nominal measurement
Grade	Overall grade given to the housing unit, based on the King County grading system(1-13)	categorical	nominal measurement
Sqft_above	square footage of house apart from the basement	integer	ratio measurement
Sqft_basement	square footage of the basement	integer	ratio measurement
Yr_built	Built Year	categorical	nominal measurement
Yr_renovated	Year when the house was renovated	categorical	nominal measurement
Zipcode	zip code	categorical	nominal measurement
Lat	Latitude coordinate	integer	interval measurement
Long	Longitude coordinate	integer	interval measurement
Sqft_living15	Living room area in 2015(implies some renovations)	integer	ratio measurement
Sqft_lot15	Lot size area in 2015(implies some renovations)	integer	ratio measurement
Median_income_by_zip	Median income from 2011-2015 American Community Survey 5-Year Estimates	integer	ratio measurement

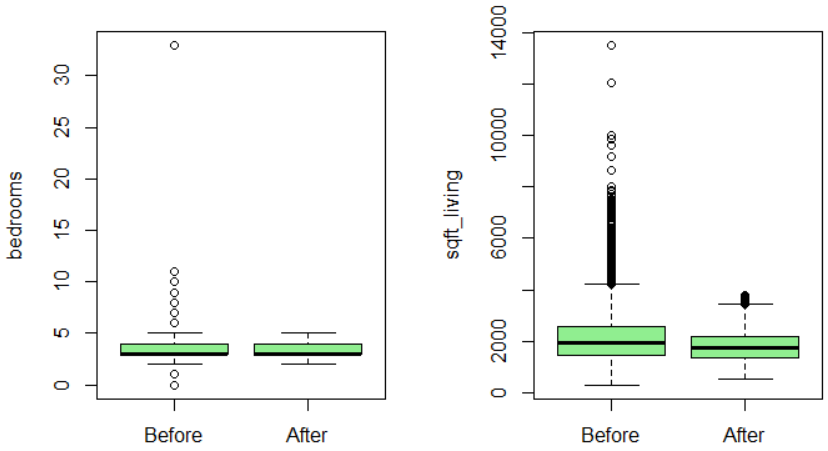
Mean_income_by_zip	Mean income from 2011-2015 American Community Survey 5-Year Estimates	integer	ratio measurement
Population_by_zip	population from 2011-2015 American Community Survey 5-Year Estimates	integer	ratio measurement
People_employed_by_zip	Civilian employed population 16 years and over from 2011-2015 American Community Survey 5-Year Estimates	integer	ratio measurement
People_educated_by_zip	Number of people who attend school from 2011-2015 American Community Survey 5-Year Estimates	integer	ratio measurement
Bachelor_degrees_by_zip	Number of people with bachelor degrees from 2011-2015 American Community Survey 5-Year Estimates	integer	ratio measurement
Grad_degrees_by_zip	Number of people with graduate or professional degrees from 2011-2015 American Community Survey 5-Year Estimates	integer	ratio measurement
Neighborhood	Houses neighborhood	categorical	nominal measurement

Table 3: Data Columns Explanations

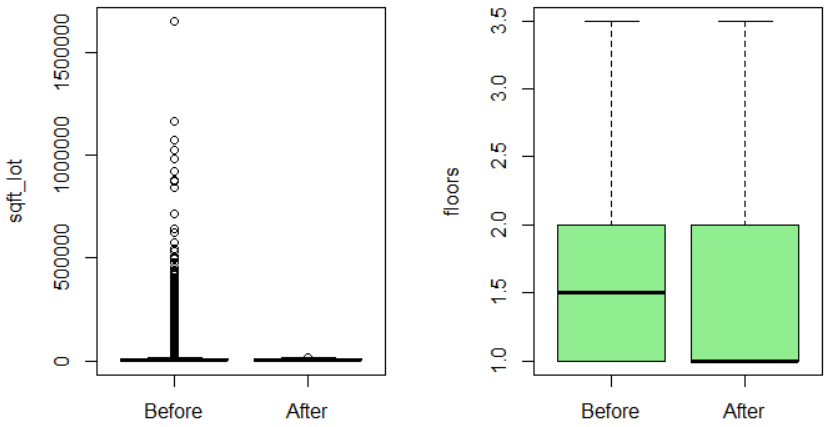
Data Cleaning



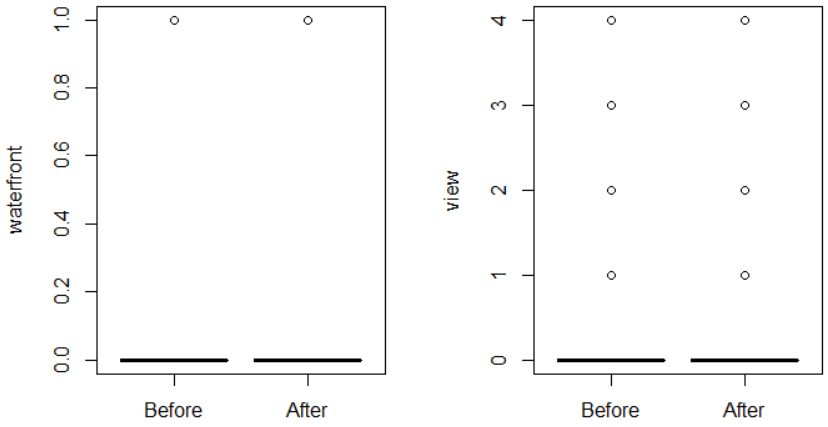
(1) Price and Bathrooms



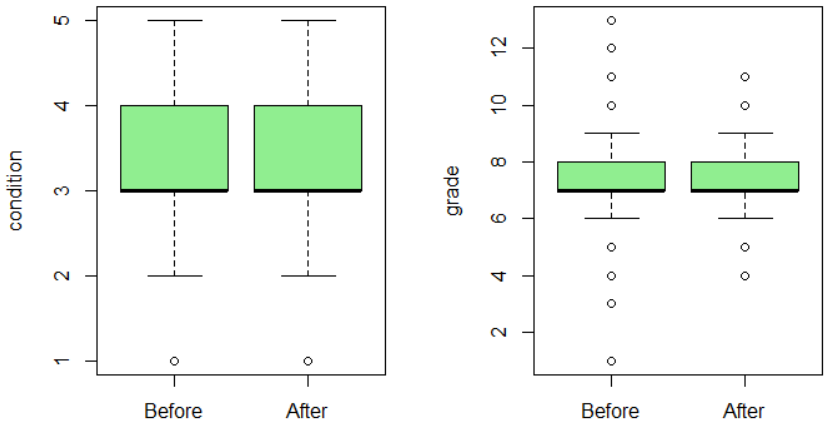
(2) Bedrooms and Sqrt of Living



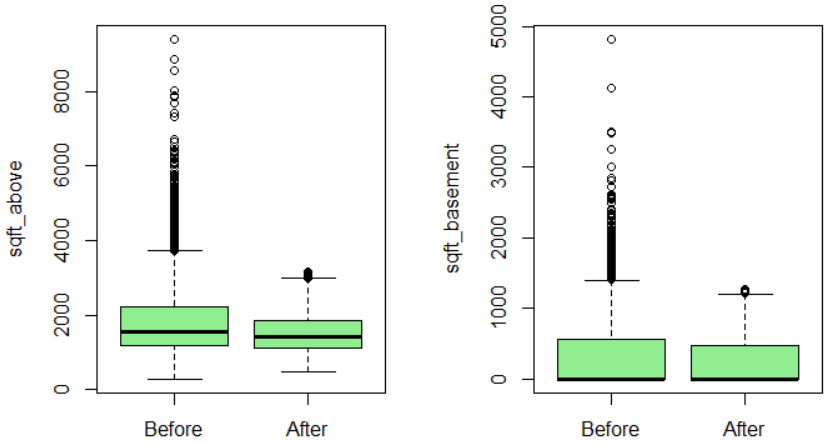
(3) Sqrt Lot and Floors



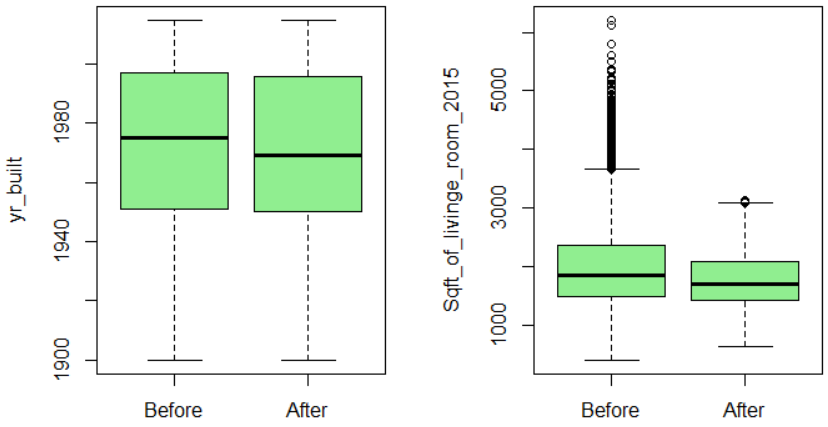
(4) Waterfront and View



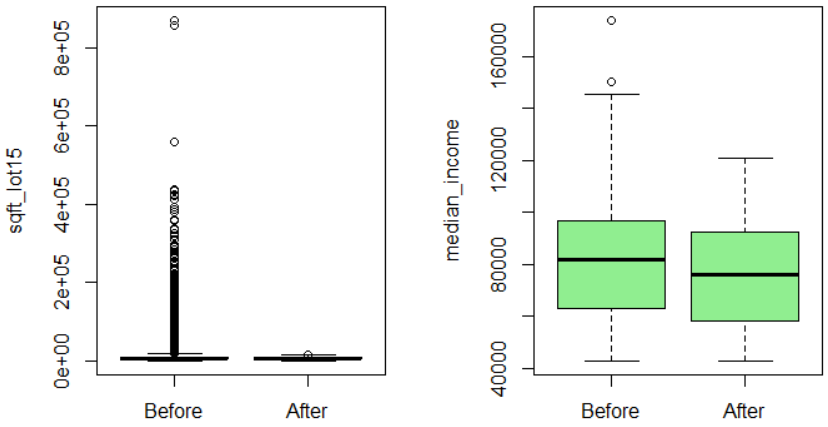
(5) Condition and Grade



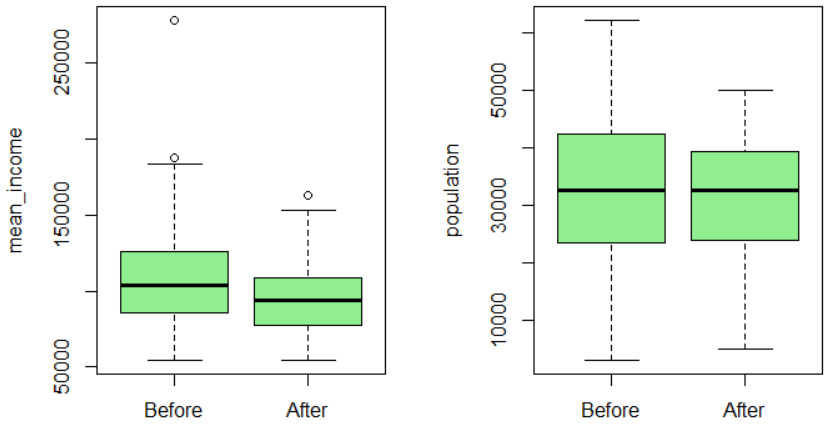
(6) Sqrt of the Above Space and Sqrt of Basement



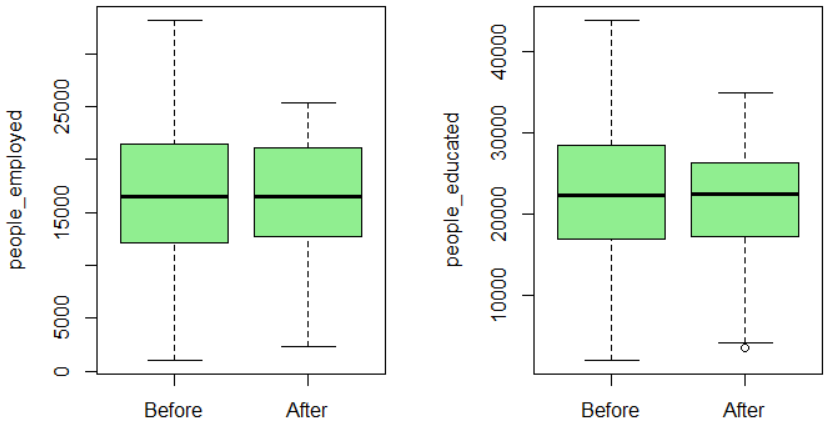
(7) The Year Built and the Sqrt of Living Room (2015)



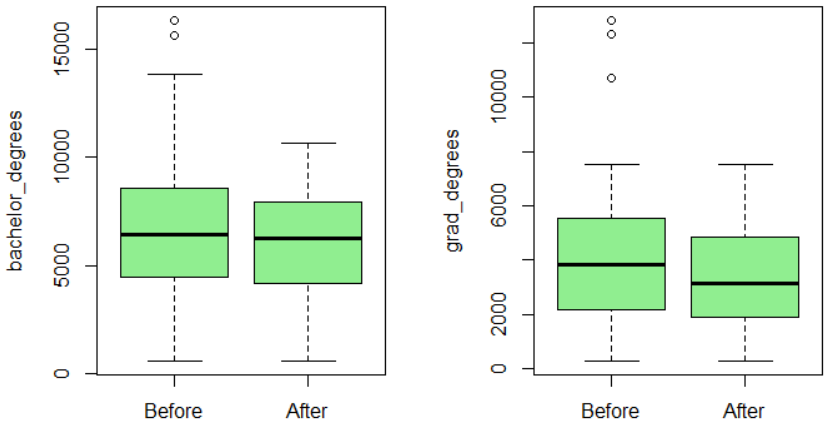
(8) Sqrt of Lot and the Median Income



(9) Mean Income and Population



(10) People Employed and People Educated



(11) The number of Bachelor Degree and the number of Graduate Degree

Figure 9: Boxplot before and after removing the outliers

Descriptive Statistics

	price	bedrooms	bathrooms	Sqft_of_home	sqft_lot
min	78000.00	2.00	0.50	520.00	520.00
max	1127000.00	5.00	4.00	3760.00	18184.00
range	1049000.00	3.00	3.50	3240.00	17664.00
sum	6021751806.00	45760.00	27670.25	25158122.00	98276196.00
median	389000.00	3.00	2.00	1720.00	7079.00
mean	425715.93	3.24	1.96	1778.59	6947.77
	floors	waterfront	view	condition	grade
min	1.00	0.00	0.00	1.00	4.00
max	3.50	1.00	4.00	5.00	11.00
range	2.50	1.00	4.00	4.00	7.00
sum	20436.50	17.00	1853.00	48321.00	103597.00
median	1.00	0.00	0.00	3.00	7.00
mean	1.44	0.00	0.13	3.42	7.32
	sqft_above	sqft_basement	yr_built	Sqft_of_livinge_room_2015	Sqft_of_lot_2015
min	480.00	0.00	1900.00	620.00	651.00
max	3172.00	1270.00	2015.00	3130.00	14733.00
range	2692.00	1270.00	115.00	2510.00	14082.00
sum	21833891.00	3324231.00	27860576.00	24987583.00	96570145.00
median	1410.00	0.00	1969.00	1690.00	7200.00
mean	1543.58	235.01	1969.64	1766.53	6827.16
	Median_Income	Mean_Income	population	people_employed	people_educated
min	42965.00	54713.00	4918.00	2341.00	3418.00
max	120970.00	162949.00	50065.00	25296.00	34944.00
range	78005.00	108236.00	45147.00	22955.00	31526.00
sum	1080840868.00	1380799919.00	451787925.00	236720601.00	316212315.00
median	75839.00	94174.00	32569.00	16520.00	22367.00
mean	76411.51	97617.53	31939.76	16735.28	22355.06
	bachelor_degrees	grad_degrees			
min	595.00	303.00			
max	10665.00	7517.00			
range	10070.00	7214.00			
sum	88296167.00	49983341.00			
median	6235.00	3142.00			
mean	6242.22	3533.64			

Table 4: Descriptive Statistics

Result for GLM (Model selected by AIC and BIC)

	price							
Predictors	*Estimates*	*std. Error*	*std. Beta*	*standardized std. Error*	*CI*	*standardized CI*	*Statistic*	*p*
(Intercept)	54239.25	36992.88			-18265.46 – 126743.96		1.47	0.143
grad degrees	50.49	1.03	0.50	0.01	48.48 – 52.51	0.48 – 0.52	49.13	**<0.001**
Sqft of home	83.15	2.77	0.26	0.01	77.72 – 88.57	0.25 – 0.28	30.02	**<0.001**
people educated	-10.71	0.16	-0.38	0.01	-11.03 – -10.40	-0.39 – -0.37	-66.75	**<0.001**
view 1	74112.14	6420.26	0.05	0.00	61528.66 – 86695.62	0.04 – 0.05	11.54	**<0.001**
view 2	74144.55	4028.81	0.07	0.00	66248.23 – 82040.87	0.07 – 0.08	18.40	**<0.001**
view 3	92267.28	6741.20	0.05	0.00	79054.78 – 105479.79	0.05 – 0.06	13.69	**<0.001**
view 4	209888.67	13274.63	0.07	0.00	183870.87 – 235906.46	0.06 – 0.08	15.81	**<0.001**
grade 5	31295.43	27662.85	0.02	0.02	-22922.76 – 85513.63	-0.01 – 0.05	1.13	0.258
grade 6	61598.45	26890.44	0.11	0.05	8894.15 – 114302.75	0.02 – 0.20	2.29	**0.022**
grade 7	96621.45	26902.79	0.26	0.07	43892.95 – 149349.95	0.12 – 0.41	3.59	**<0.001**
grade 8	144675.61	26991.99	0.36	0.07	91772.28 – 197578.94	0.23 – 0.49	5.36	**<0.001**
grade 9	226503.03	27173.35	0.33	0.04	173244.24 – 279761.82	0.25 – 0.41	8.34	**<0.001**
grade 10	304959.95	28133.35	0.16	0.01	249819.60 – 360100.31	0.13 – 0.19	10.84	**<0.001**
grade 11	522428.34	65348.87	0.03	0.00	394346.90 – 650509.78	0.03 – 0.04	7.99	**<0.001**
condition 2	63716.09	24191.34	0.03	0.01	16301.92 – 111130.25	0.01 – 0.05	2.63	**0.008**
condition 3	93053.37	22772.65	0.24	0.06	48419.79 – 137686.95	0.13 – 0.36	4.09	**<0.001**
condition 4	109699.44	22777.68	0.26	0.05	65056.01 – 154342.88	0.16 – 0.37	4.82	**<0.001**
condition 5	132984.40	22863.00	0.20	0.03	88173.74 – 177795.05	0.13 – 0.26	5.82	**<0.001**
bachelor degrees	10.82	0.90	0.14	0.01	9.06 – 12.57	0.12 – 0.16	12.08	**<0.001**
yr built 1901	-58019.59	21051.31	-0.01	0.00	-99279.40 – -16759.79	-0.02 – -0.00	-2.76	**0.006**
yr built 1902	-28215.99	21770.52	-0.01	0.00	-70885.42 – 14453.45	-0.01 – 0.00	-1.30	0.195
yr built 1903	-48655.92	17544.72	-0.01	0.00	-83042.94 – -14268.90	-0.02 – -0.00	-2.77	**0.006**
yr built 1904	-9749.52	17413.19	-0.00	0.00	-43878.74 – 24379.71	-0.01 – 0.01	-0.56	0.576
yr built 1905	14514.95	17152.75	0.00	0.01	-19103.82 – 48133.73	-0.01 – 0.01	0.85	0.397
yr built 1906	-22659.94	15986.02	-0.01	0.01	-53991.97 – 8672.09	-0.02 – 0.00	-1.42	0.156
yr built 1907	18861.71	16226.84	0.01	0.01	-12942.31 – 50665.74	-0.00 – 0.02	1.16	0.245
yr built 1908	-52364.90	15997.64	-0.02	0.01	-83719.71 – -21010.10	-0.03 – -0.01	-3.27	**0.001**
yr built 1909	-14229.46	15151.27	-0.01	0.01	-43925.39 – 15466.48	-0.02 – 0.01	-0.94	0.348
yr built 1910	-9106.53	14130.00	-0.00	0.01	-36800.81 – 18587.76	-0.02 – 0.01	-0.64	0.519

yr built 1911	-5806.53	17293.23	-0.00	0.01	-39700.64 – 28087.57	-0.01 – 0.01	-0.34	0.737
yr built 1912	-1375.52	15683.69	-0.00	0.01	-32114.99 – 29363.96	-0.01 – 0.01	-0.09	0.930
yr built 1913	10805.88	17541.94	0.00	0.00	-23575.69 – 45187.44	-0.01 – 0.01	0.62	0.538
yr built 1914	-8709.33	17674.32	-0.00	0.00	-43350.36 – 25931.71	-0.01 – 0.01	-0.49	0.622
yr built 1915	-7357.34	16819.80	-0.00	0.01	-40323.54 – 25608.86	-0.01 – 0.01	-0.44	0.662
yr built 1916	-7402.18	16075.35	-0.00	0.01	-38909.29 – 24104.93	-0.01 – 0.01	-0.46	0.645
yr built 1917	-22396.16	17279.52	-0.01	0.01	-56263.40 – 11471.08	-0.02 – 0.00	-1.30	0.195
yr built 1918	-35828.68	14311.90	-0.02	0.01	-63879.48 – -7777.88	-0.03 – -0.00	-2.50	**0.012**
yr built 1919	12982.28	15317.48	0.00	0.01	-17039.43 – 43003.99	-0.01 – 0.02	0.85	0.397
yr built 1920	-9537.50	15145.15	-0.00	0.01	-39221.44 – 20146.44	-0.01 – 0.01	-0.63	0.529
yr built 1921	-27110.41	16416.69	-0.01	0.01	-59286.54 – 5065.71	-0.02 – 0.00	-1.65	0.099
yr built 1922	-4262.09	15146.47	-0.00	0.01	-33948.61 – 25424.44	-0.01 – 0.01	-0.28	0.778
yr built 1923	-3311.58	15922.91	-0.00	0.01	-34519.90 – 27896.74	-0.01 – 0.01	-0.21	0.835
yr built 1924	-3394.31	14217.62	-0.00	0.01	-31260.33 – 24471.71	-0.01 – 0.01	-0.24	0.811
yr built 1925	-8234.85	13660.09	-0.00	0.01	-35008.13 – 18538.43	-0.02 – 0.01	-0.60	0.547
yr built 1926	-5681.36	13694.13	-0.00	0.01	-32521.37 – 21158.65	-0.02 – 0.01	-0.41	0.678
yr built 1927	-15705.27	14869.99	-0.01	0.01	-44849.91 – 13439.37	-0.02 – 0.01	-1.06	0.291
yr built 1928	-22665.09	14290.61	-0.01	0.01	-50674.17 – 5343.99	-0.02 – 0.00	-1.59	0.113
yr built 1929	-21266.22	14257.31	-0.01	0.01	-49210.04 – 6677.60	-0.02 – 0.00	-1.49	0.136
yr built 1930	-41897.28	15775.16	-0.01	0.01	-72816.02 – -10978.54	-0.02 – -0.00	-2.66	**0.008**
yr built 1931	-22571.55	17554.04	-0.01	0.00	-56976.83 – 11833.73	-0.02 – 0.00	-1.29	0.199
yr built 1932	18602.93	21141.48	0.00	0.00	-22833.62 – 60039.48	-0.00 – 0.01	0.88	0.379
yr built 1933	-1669.38	25020.24	-0.00	0.00	-50708.15 – 47369.39	-0.01 – 0.01	-0.07	0.947
yr built 1934	-18397.27	28833.07	-0.00	0.00	-74909.05 – 38114.51	-0.01 – 0.01	-0.64	0.523
yr built 1935	-21789.32	28700.31	-0.00	0.00	-78040.90 – 34462.26	-0.01 – 0.01	-0.76	0.448
yr built 1936	-8691.78	19836.93	-0.00	0.00	-47571.45 – 30187.90	-0.01 – 0.01	-0.44	0.661
yr built 1937	-46646.12	18899.09	-0.01	0.00	-83687.65 – -9604.59	-0.02 – -0.00	-2.47	**0.014**
yr built 1938	-2893.47	19336.61	-0.00	0.00	-40792.53 – 35005.58	-0.01 – 0.01	-0.15	0.881
yr built 1939	-11013.36	15018.54	-0.00	0.01	-40449.16 – 18422.43	-0.02 – 0.01	-0.73	0.463
yr built 1940	-20107.83	13694.25	-0.01	0.01	-46948.06 – 6732.40	-0.02 – 0.00	-1.47	0.142
yr built 1941	-14410.19	13628.41	-0.01	0.01	-41121.39 – 12301.01	-0.02 – 0.01	-1.06	0.290
yr built 1942	-33781.74	12770.56	-0.02	0.01	-58811.59 – -8751.90	-0.04 – -0.01	-2.65	**0.008**

yr built 1943	-52135.63	13178.83	-0.03	0.01	-77965.67 – -26305.58	-0.04 – -0.01	-3.96	**<0.001**
yr built 1944	-47199.57	13382.77	-0.02	0.01	-73429.31 – -20969.82	-0.04 – -0.01	-3.53	**<0.001**
yr built 1945	-47650.54	15447.07	-0.02	0.01	-77926.25 – -17374.84	-0.03 – -0.01	-3.08	**0.002**
yr built 1946	-31960.26	14412.50	-0.01	0.01	-60208.25 – -3712.27	-0.03 – -0.00	-2.22	**0.027**
yr built 1947	-49657.70	12652.62	-0.03	0.01	-74456.37 – -24859.02	-0.05 – -0.02	-3.92	**<0.001**
yr built 1948	-46326.20	12819.02	-0.03	0.01	-71451.02 – -21201.39	-0.04 – -0.01	-3.61	**<0.001**
yr built 1949	-31128.09	13171.96	-0.02	0.01	-56944.67 – -5311.52	-0.03 – -0.00	-2.36	**0.018**
yr built 1950	-44143.63	12806.54	-0.03	0.01	-69243.98 – -19043.27	-0.04 – -0.01	-3.45	**0.001**
yr built 1951	-47145.13	12979.96	-0.03	0.01	-72585.38 – -21704.89	-0.04 – -0.01	-3.63	**<0.001**
yr built 1952	-40334.32	12971.21	-0.02	0.01	-65757.44 – -14911.21	-0.04 – -0.01	-3.11	**0.002**
yr built 1953	-39867.84	12994.33	-0.02	0.01	-65336.26 – -14399.42	-0.04 – -0.01	-3.07	**0.002**
yr built 1954	-56333.78	12398.90	-0.04	0.01	-80635.18 – -32032.39	-0.06 – -0.02	-4.54	**<0.001**
yr built 1955	-74467.08	12575.47	-0.05	0.01	-99114.54 – -49819.61	-0.07 – -0.03	-5.92	**<0.001**
yr built 1956	-61987.94	12962.61	-0.04	0.01	-87394.20 – -36581.69	-0.05 – -0.02	-4.78	**<0.001**
yr built 1957	-61143.95	13142.41	-0.03	0.01	-86902.60 – -35385.30	-0.05 – -0.02	-4.65	**<0.001**
yr built 1958	-55217.68	13121.98	-0.03	0.01	-80936.29 – -29499.07	-0.05 – -0.02	-4.21	**<0.001**
yr built 1959	-69205.32	12404.61	-0.05	0.01	-93517.90 – -44892.74	-0.07 – -0.03	-5.58	**<0.001**
yr built 1960	-64850.35	12891.61	-0.04	0.01	-90117.45 – -39583.26	-0.05 – -0.02	-5.03	**<0.001**
yr built 1961	-76726.60	12884.39	-0.05	0.01	-101979.54 – -51473.65	-0.06 – -0.03	-5.96	**<0.001**
yr built 1962	-74900.01	12526.06	-0.05	0.01	-99450.63 – -50349.39	-0.07 – -0.04	-5.98	**<0.001**
yr built 1963	-74878.60	12904.13	-0.05	0.01	-100170.23 – -49586.96	-0.06 – -0.03	-5.80	**<0.001**
yr built 1964	-73727.96	13807.76	-0.04	0.01	-100790.68 – -46665.24	-0.05 – -0.02	-5.34	**<0.001**
yr built 1965	-84553.13	13433.88	-0.04	0.01	-110883.06 – -58223.20	-0.06 – -0.03	-6.29	**<0.001**
yr built 1966	-85851.83	12850.00	-0.05	0.01	-111037.36 – -60666.30	-0.07 – -0.04	-6.68	**<0.001**
yr built 1967	-74295.77	12465.22	-0.05	0.01	-98727.15 – -49864.40	-0.07 – -0.04	-5.96	**<0.001**
yr built 1968	-71502.96	12453.16	-0.05	0.01	-95910.71 – -47095.21	-0.07 – -0.03	-5.74	**<0.001**
yr built 1969	-87191.67	12641.55	-0.06	0.01	-111968.66 – -62414.68	-0.07 – -0.04	-6.90	**<0.001**
yr built 1970	-79385.50	14614.36	-0.03	0.01	-108029.12 – -50741.88	-0.04 – -0.02	-5.43	**<0.001**
yr built 1971	-80836.91	15086.72	-0.03	0.01	-110406.34 – -51267.48	-0.04 – -0.02	-5.36	**<0.001**
yr built 1972	-91847.38	14171.27	-0.04	0.01	-119622.56 – -64072.20	-0.05 – -0.03	-6.48	**<0.001**
yr built 1973	-83548.10	14376.53	-0.04	0.01	-111725.57 – -55370.63	-0.05 – -0.02	-5.81	**<0.001**
yr built 1974	-87733.28	14029.87	-0.04	0.01	-115231.32 – -60235.23	-0.05 – -0.03	-6.25	**<0.001**

yr built 1975	-97789.90	13811.49	-0.05	0.01	-124859.92 – -70719.88	-0.06 – -0.03	-7.08	**<0.001**
yr built 1976	-86737.56	13328.96	-0.05	0.01	-112861.85 – -60613.28	-0.06 – -0.03	-6.51	**<0.001**
yr built 1977	-88258.97	12484.55	-0.06	0.01	-112728.24 – -63789.71	-0.08 – -0.05	-7.07	**<0.001**
yr built 1978	-105691.81	12574.80	-0.07	0.01	-130337.97 – -81045.65	-0.09 – -0.06	-8.41	**<0.001**
yr built 1979	-98430.82	12634.53	-0.07	0.01	-123194.04 – -73667.60	-0.08 – -0.05	-7.79	**<0.001**
yr built 1980	-79586.64	13398.62	-0.04	0.01	-105847.45 – -53325.84	-0.06 – -0.03	-5.94	**<0.001**
yr built 1981	-73456.65	13606.68	-0.04	0.01	-100125.26 – -46788.04	-0.05 – -0.02	-5.40	**<0.001**
yr built 1982	-86562.31	16288.17	-0.03	0.01	-118486.53 – -54638.08	-0.04 – -0.02	-5.31	**<0.001**
yr built 1983	-75682.26	13707.00	-0.04	0.01	-102547.50 – -48817.02	-0.05 – -0.02	-5.52	**<0.001**
yr built 1984	-65437.65	13996.05	-0.03	0.01	-92869.39 – -38005.90	-0.04 – -0.02	-4.68	**<0.001**
yr built 1985	-81298.62	13436.24	-0.04	0.01	-107633.16 – -54964.07	-0.06 – -0.03	-6.05	**<0.001**
yr built 1986	-89792.62	13440.57	-0.05	0.01	-116135.65 – -63449.59	-0.06 – -0.03	-6.68	**<0.001**
yr built 1987	-97526.59	13017.68	-0.06	0.01	-123040.77 – -72012.41	-0.07 – -0.04	-7.49	**<0.001**
yr built 1988	-109130.54	13137.21	-0.06	0.01	-134878.99 – -83382.09	-0.08 – -0.05	-8.31	**<0.001**
yr built 1989	-108166.99	13062.41	-0.06	0.01	-133768.84 – -82565.14	-0.08 – -0.05	-8.28	**<0.001**
vr built 1990	-117717.35	12924.63	-0.07	0.01	-143049.16 – -92385.53	-0.09 – -0.06	-9.11	**<0.001**
yr built 1991	-107148.61	13667.99	-0.05	0.01	-133937.39 – -80359.84	-0.07 – -0.04	-7.84	**<0.001**
yr built 1992	-105697.03	13563.75	-0.05	0.01	-132281.49 – -79112.56	-0.07 – -0.04	-7.79	**<0.001**
yr built 1993	-99908.11	13451.80	-0.05	0.01	-126273.15 – -73543.06	-0.07 – -0.04	-7.43	**<0.001**
yr built 1994	-98815.35	13003.02	-0.06	0.01	-124300.80 – -73329.89	-0.07 – -0.04	-7.60	**<0.001**
yr built 1995	-87850.47	14303.70	-0.04	0.01	-115885.21 – -59815.74	-0.05 – -0.03	-6.14	**<0.001**
yr built 1996	-86310.74	13859.71	-0.04	0.01	-113475.26 – -59146.22	-0.05 – -0.03	-6.23	**<0.001**
yr built 1997	-102208.52	14255.13	-0.05	0.01	-130148.06 – -74268.99	-0.06 – -0.03	-7.17	**<0.001**
yr built 1998	-109896.30	13471.23	-0.06	0.01	-136299.43 – -83493.17	-0.07 – -0.04	-8.16	**<0.001**
yr built 1999	-106957.02	13334.32	-0.06	0.01	-133091.81 – -80822.24	-0.07 – -0.04	-8.02	**<0.001**
yr built 2000	-94185.55	13881.74	-0.05	0.01	-121393.26 – -66977.84	-0.06 – -0.03	-6.78	**<0.001**
yr built 2001	-109973.05	13095.85	-0.06	0.01	-135640.44 – -84305.65	-0.08 – -0.05	-8.40	**<0.001**
yr built 2002	-111531.07	13340.25	-0.06	0.01	-137677.48 – -85384.66	-0.07 – -0.05	-8.36	**<0.001**
yr built 2003	-107014.62	12305.87	-0.08	0.01	-131133.68 – -82895.56	-0.10 – -0.06	-8.70	**<0.001**
yr built 2004	-116330.22	12398.64	-0.09	0.01	-140631.11 – -92029.34	-0.10 – -0.07	-9.38	**<0.001**
yr built 2005	-127281.40	12332.59	-0.10	0.01	-151452.83 – -103109.97	-0.12 – -0.08	-10.32	**<0.001**
yr built 2006	-120167.93	12394.32	-0.09	0.01	-144460.34 – -95875.52	-0.11 – -0.07	-9.70	**<0.001**

yr built 2007	-106001.82	12458.22	-0.08	0.01	-130419.48 – -81584.17	-0.10 – -0.06	-8.51	**<0.001**
yr built 2008	-104879.02	12627.61	-0.07	0.01	-129628.68 – -80129.37	-0.09 – -0.06	-8.31	**<0.001**
yr built 2009	-96253.73	12864.05	-0.06	0.01	-121466.80 – -71040.65	-0.08 – -0.04	-7.48	**<0.001**
yr built 2010	-95057.32	13808.73	-0.05	0.01	-122121.94 – -67992.70	-0.06 – -0.03	-6.88	**<0.001**
yr built 2011	-99355.13	14144.06	-0.05	0.01	-127076.99 – -71633.27	-0.06 – -0.03	-7.02	**<0.001**
yr built 2012	-93663.86	13511.67	-0.05	0.01	-120146.25 – -67181.47	-0.06 – -0.04	-6.93	**<0.001**
yr built 2013	-76326.10	13374.38	-0.04	0.01	-102539.40 – -50112.80	-0.06 – -0.03	-5.71	**<0.001**
yr built 2014	-77324.62	12200.26	-0.07	0.01	-101236.68 – -53412.55	-0.09 – -0.05	-6.34	**<0.001**
yr built 2015	-67219.49	21223.39	-0.01	0.00	-108816.58 – -25622.40	-0.02 – -0.01	-3.17	**0.002**
bathrooms	17306.71	1883.13	0.06	0.01	13615.83 – 20997.58	0.05 – 0.08	9.19	**<0.001**
Sqft of livinge room 2015	16.98	2.31	0.04	0.01	12.46 – 21.50	0.03 – 0.06	7.36	**<0.001**
waterfront 1	154289.53	24293.59	0.03	0.00	106674.97 – 201904.09	0.02 – 0.04	6.35	**<0.001**
sqft above	15.83	2.88	0.05	0.01	10.19 – 21.48	0.03 – 0.07	5.50	**<0.001**
Sqft of lot 2015	-1.69	0.37	-0.03	0.01	-2.43 – -0.96	-0.04 – -0.01	-4.51	**<0.001**
floors	-7538.70	2325.87	-0.02	0.01	-12097.32 – -2980.08	-0.03 – -0.01	-3.24	**0.001**
Observations	14145							

Figure 14: Summary Data for general linear regression model based on the AIC and BIC result

VIII Reference

[1] *AIC vs. BIC*, Penn State University, college of Health and Human Development. Retrieved from : https://methodology.psu.edu/AIC-vs-BIC

[2] *AMERICAN FactFinder*, United States Census Bureau. Retrieved from: https://factfinder.census.gov/faces/nav/jsf/pages/index.xhtml

[3] *Anisotropic Feature Ranking RFE, Random Forest, linear models*, Arthur Tok, Kaggle. Retrieved from : https://www.kaggle.com/arthurtok/feature-ranking-rfe-random-forest-linear-models/notebook

[4] *Discovering the Hidden Structure of House Prices with a Non-Parametric Latent Manifold Model*, Sumit Chopra, Courant Institute of Mathematical Sciences, New York University. Retrieved from : http://yann.lecun.com/exdb/publis/pdf/chopra-kdd-07.pdf

[5] *House Sales Visualization*,Emirhan Ergin, Kaggle. Retrieved from : https://www.kaggle.com/darkcore/house-sales-visualization/notebook

[6] *Hyperparameter Tuning the Random Forest in Python*, William Koehrsen, Towards Data Science. Retrieved from: https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

[7] *Machine Learning with XGBoost*, Scott Lundberg. Retrieved from : https://towardsdatascience.com/interpretable-machine-learning-with-xgboost-9ec80d148d27

[8] *Metro Neighborhoods in King County*, KCGIS Center, King County Gov. Retrieved from: http://www5.kingcounty.gov/sdc/Metadata.aspx?Layer=neighborhood

[9] *Multiple Linear Regression*, Divan, Kaggle. Retrieved from : https://www.kaggle.com/divan0/multiple-linear-regression/notebook

[10] *Multicollinearity in Regression Analysis: Problems, Detection, and Solutions*, Jim Frost. Retrieved from : http://statisticsbyjim.com/regression/multicollinearity-in-regression-analysis/

[11] *Predicting House Prices*, Burhan Y. Kiyakoglu, Kaggle. Retrieved from: www.kaggle.com/burhanykiyakoglu/predicting-house-prices/notebook

[12] *Predicting House Prices Using Multiple Listings Data*, R.A. Dubin, The Journal of Real Estate Finance and Economics (1998) 17: 35. Retrieved from : https://doi.org/10.1023/A:1007751112669

[13] *Random Forest in Python*, William Koehrsen, Towards Data Science. Retrieved from: https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

[14] *Shiller Home Price Index*, S&P Dow Jones Indices. Retrieved from : https://us.spindices.com/index-family/real-estate/sp-corelogic-case-shiller

[15] *Turi Machine Learning Platform User Guide*, turi.com. Retrieved from: https://turi.com/learn/userguide/supervised-learning/random_forest_regression.html

[16] *XGBoost Parameters*. Retrieved from : https://xgboost.readthedocs.io/en/latest/parameter.html

[17] *XGBoost: A Scalable Tree Boosting System*,Tianqi Chen,Carlos Guestrin . Retrieved from : https://www.kdd.org/kdd2016/papers/files/rfp0697-chenAemb.pdf