ATLANTIS PRESS
Advances in Economics, Business and Management Research, volume 211
Proceedings of the 2022 7th International Conference on Financial Innovation and Economic Development (ICFIED 2022)

House Price Prediction Based on Machine Learning: A Case of King County

Yijia Wang[1, †] and Qiaotong Zhao[2, *, †]

[1] *Queen's University. 39-2250 Rockingham Drive, L6H 6J3, Oakville, ON, Canada. Email: 18yw148@queensu.ca*
[2] *Civil Aviation University of China. No.2898, Jinbei Highway, Dongli District, Tianjin, China 300300. Email: zhaoqiaotong@163.com*
**Corresponding author. Email: zhaoqiaotong@163.com*
[†]*These authors contributed equally.*

ABSTRACT

This paper focuses on formulating a feasible method for house price prediction. A dataset containing features and house price of King County in the US is used. During the data preprocessing, extreme values are winsorized and highly correlated features are removed. Eight models including Catboost, lightGBM and XGBoost serve as candidate models. They are evaluated by several evaluation indicators, including rooted mean square error, R-squared score, adjusted R-squared score and K-fold cross validation score. The model that has low RMSE, achieves a high R-squared score and adjusted R-squared score, especially in the test set, and acquires a high score in cross validation is considered a better model. This paper finds out that Catboost performs the best among all models and can be used for house price prediction. Location, living space and condition of the house are the most important features influencing house price. After comparison and contrast with other papers, it is attested that findings in this paper conform to real life. This paper formulates a model that fits better than preceding studies for house price prediction and makes necessary supplement to the exploration of features that influence house price from a microscope.

Keywords: *Catboost, House Price, King County, Prediction.*

1. INTRODUCTION

The trend of house prices is always a controversial topic as its fluctuation will pose a huge effect on the entire economy. The rise in house price means growth in non-financial assets which ultimately increase personal wealth, stimulating household consumption and boosting the economy; however, a decrease in house price limits an individual's borrowing capacity, crowding out investments due to the evaporation in the value of collaterals [1]. The shock in the global economy caused by the 2008 housing bubble perfectly explains the importance of a stable and measurable house price. The turmoil in house prices causes an unexpected rise in real long-term interest rates, bankruptcy in financial institutions and global economic depression [2]. Although it is hard to control the house price, it is possible to predict it.

Many scholars have conducted research on this issue. For instance, Hirata et al. have used time-series models to determine that house prices have become more synchronized over time and the FAVAR model to find out that global interest rate shock has the most considerable influence on global house price, especially in the US [3]. Shishir Mathur has provided insight from a micro perspective in his report, stating that quality and size are two factors, contributing to house price [4]. In his opinion, this assumption can be explained through the perceived value of the house. The property assessors will evaluate the size and quality of the house during value assessment processes for house reselling, which will determine the value of the house. Property developers also take houses' quality and size into consideration while they initially design the project and pricing for the property. A bigger size and better quality will bring a higher perceived value to both assessors, developers, and buyers. In Shishir Mathur's report, he also mentioned another contributor – the level of maintenance. With the increased investment in refurbishing before offering for sale, the house owners will expect a higher dealing price due to their value addition through the maintenance. Although prior researches make a valid analysis, they failed to discuss the simultaneous effect of those factors.

Some factors may contribute more to the results than others. Besides, their conclusions are based on theoretical knowledge and lack practical proof.

This paper goes beyond previous economic analysis and uses machine learning to explore the country-wide house price. This paper also assumes that the two factors, size and quality mentioned by Shishir Mathur, will affect the house price, but this paper will use Machine learning algorithms to prove the relationship. Other than these two factors, some other factors on house price, including location, size, and overall structure of the house and grading from the agency will also be evaluated. Machine learning's most obvious advantages are that it can automatically solve a wide range of problems and efficiently handle big datasets [5]. These two benefits allow us to prove the assumptions through analyzing a huge amount of historical data and taking multiple factors into consideration to present a comprehensive model efficiently.

The research studies the house price in King County, US, during a 2-year period from 2014 to 2015. According to the data gathered by Washington Government, King County has the highest estimated population of 2,052,800 in 2015 among all counties in Washington [6]. With a higher population, King County has more potential house buyers and higher house demands; thus, house price data in King County will be more complete and more precise. This paper employs different technical models, including Catboost, LightGBM, XGBoost, Random Forest and regressions to identify the important influencer of the house price. The best model will be selected through training and testing, which will allow us to have the most accurate result. The results will conclude important micro features in determining the house price, including the location, size, and gradings. This will provide a guide for future house price prediction to not only consider the macro effects but also think about the micro factors.

The rest of the paper is organized as follows: section 2 and section 3 introduce the source of the data and the methodology used in the evaluation. The fourth section discusses the evaluation and the results. Eventually, the conclusions are in section 5.

2. DATA

2.1. DESCRIPTION OF THE DATASET

The dataset used to predict the sales price of houses in King County comes from Kaggle. It includes 21613 observations of 20 house features and one house price column for homes sold between May 2014 and May 2015.

Among the 20 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. One thing that should be mentioned is that values in the attribute, "yr_renovated", will be replaced by the difference between the year of renovation and the year sold out. Additionally, there are two evaluation scores: "Grade" and "Condition". These two attributes grade the overall condition of the house based on different scales and standards [7].

2.2. DATA PREPROCESSING

A classic phrase in computing says "garbage in, garbage out" [8]. In another word, "good" data is the origins of high-quality analysis and project design. Foxwell concluded four causes of data error: creation and pre-collection errors, collection errors, post-collection and analysis errors, and recording errors [9]. During this stage, this paper will focus on addressing collection errors of the dataset, especially the two frequently mentioned collection errors: missing variables and outliers. Since the counts of each feature are equal to the total number of observations, the data does not have any missing variables. However, there is no outliers in the data. Comparing the price at 75% quantile ($645000) with the maximum number ($7700000), feature "price" should have some outliers (Figure 1). After further analyzing the distribution of "price", the right-skewed distribution with a fat tail confirms the outliers. Thus, the price which are greater than 99th percentile number will be replaced with 99th percentile numbers. The same method will be applied in the other numerical features.

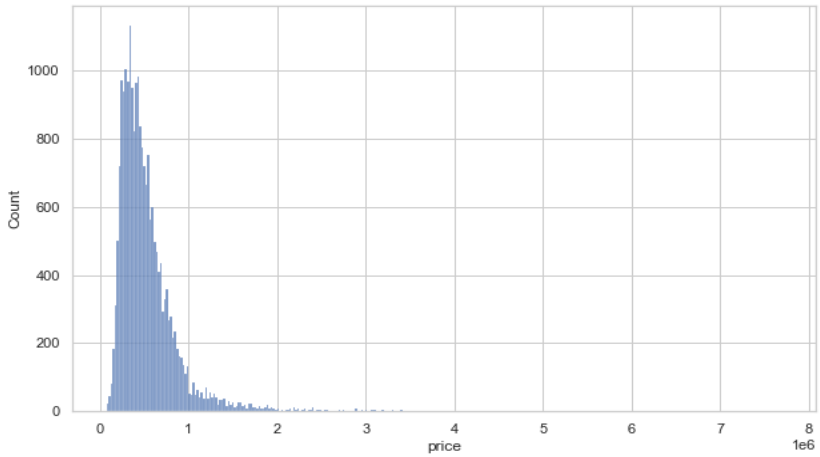


Figure 1 Distribution of the price feature.

After adjusting the dataset into a "good" version, the paper further explores the implicit meaning of the raw data. Firstly, the distribution of the features illustrates that most of the house prices collected in the data come from houses with 2 to 5 bedrooms (Figure 2), 1 or 2.5 bathrooms (Figure 3), 1 or 2 floors (Figure 4), and without waterfront and view (Figure 5). Houses with this structure will be more attractive in the real estate market. Secondly, most of them have the overall condition of a house is at the 3rd level on a scale of 1 to 5 and the overall grade given to the housing unit, based on King County grading system is at level 7 on a scale of 1 to 11. This demonstrates that the property in the data is mid-level houses.

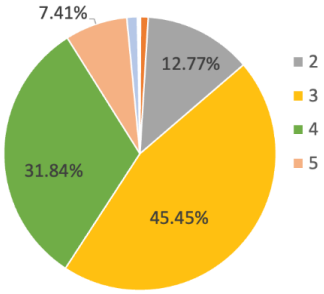


Figure 2 Number of bedrooms.

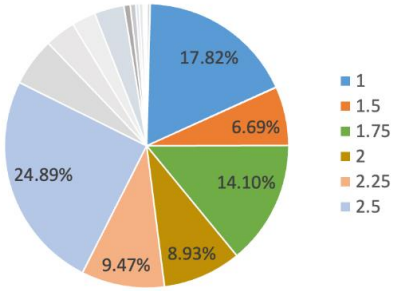


Figure 3 Number of bathrooms.

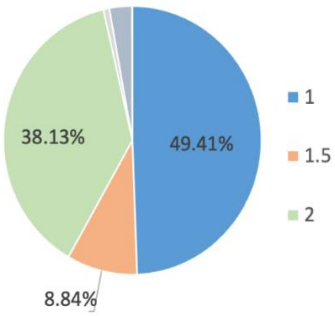


Figure 4 Number of floors.

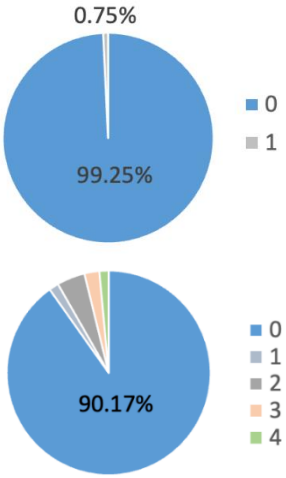
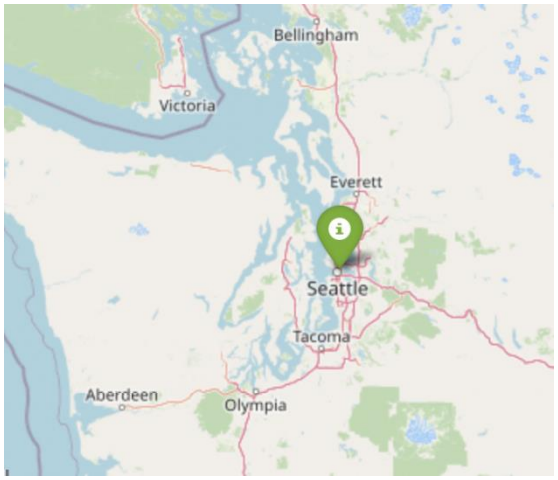
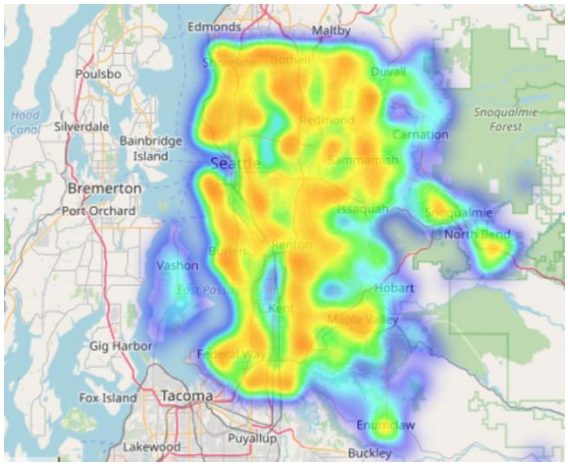


Figure 5 Number of waterfronts & number of views.

The observation from the heatmap in Figure 6, demonstrates that most of the 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. This is because most of the area in east King County are covered by forest.



a) Location of the King County.



b) The geographical distribution of the price.

Figure 6 Heatmap of house price.

2.3. *Feature Selection*

According to the correlation matrix (Figure 7), "sqft_living15" has a high correlation with "bathrooms" (0.57), "sqft_living" (0.76), "grade" (0.71), and "sqft_above" (0.73). Therefore, this feature will be dropped to avoid multi-correlation and increase the accuracy of the result. Additionally, the feature, "sqft_lot15" which has a similar meaning to "sqft_living15" will also be dropped. Lastly, because "id" does not have any noticeable relationship with house price, it will be deleted.

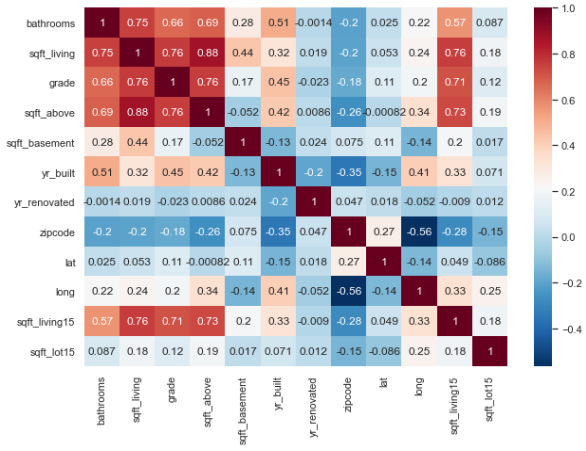


Figure 7 Correlation matrix among features.

3. METHODOLOGIES

This paper tests eight regression models, as implemented in the Sci-kit learn, XGBoost, Catboost and LightGBM package of Python. Those models include multiple linear regression, polynomial regression, lasso regression, ridge regression, random forest regression, XGBoost regression, LightGBM regression as well as Catboost regression.

3.1. *Models*

3.1.1. *Multiple linear regression and polynomial regression*

In multiple linear regression, the output is subject to $x_1, x_2, \ldots x_n$. It is determined when $\theta_0, \theta_1, \ldots \theta_n$, are chosen [10]. It can be represented as:

$$f(X) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \cdots + \theta_n x_n \quad (1)$$

where $X = [x_1, x_2, \ldots x_n]$.

Similarly, the polynomial regression model can be written as:

$$f(X) = \theta_0 + \theta_1 {x_1}^1 + \theta_2 {x_2}^2 + \cdots + \theta_n {x_n}^i \quad (2)$$

where i denotes the degree of independent variables [11].

3.1.2. *Ridge and lasso regression*

In ridge regression, the goal is to optimize the following program:

$$J(\theta) = MSE(\theta) + \alpha \frac{1}{2} \sum_{i=1}^{n} {\theta_i}^2 \quad (3)$$

where α is a parameter used to balance the regularization factor and the error. The second factor is l_2 regularization, which avoids overfitting [12].

Similarly, in lasso regression, the goal is also to optimize a program.

$$J(\theta) = MSE(\theta) + \alpha \sum_{i=1}^{n} |\theta_i| \quad (4)$$

The second factor is l_1 regularization.

3.1.3. RANDOM FOREST REGRESSION

Random forest searches for the best feature among a random set of features. It trains the model for T rounds. The best feature in each random subset is used to split the node and the combination of them generates the strong learner F(x).

3.1.4. XGBOOST AND LIGHTGBM REGRESSION

XGBoost is a method, originated from gradient boosting decision tree (GBDT). Its objective is to find the function f(x) to fit the residual error in the last node so that the loss function is reduced to the minimum [13]. LightGBM is another method based on GBDT. It adopts the histogram algorithm, which reduces the computational cost. It also adopts leaf-wise tree growth. Every time the tree grows, it splits from the node that performs the best and iterates this process. What's more, it controls the maximum depth to avoid overfitting [14].

3.1.5. CATBOOST REGRESSION

The model that outperforms other ones is Catboost, and it is worth diving deeper into its algorithm. Catboost adopts the gradient boosting procedure. Ft is built iteratively in a greedy fashion, representing a sequence of approximations. Ft is obtained from the following equation:

$$F^t = F^{t-1} + \alpha h^t \tag{5}$$

where α is the step size, and h^t is chosen to minimize the following loss function.

$$h^t = L\,(F^{t-1} + h) = EL(y, F^{t-1}(x) + h(x)) \tag{6}$$

The solution to the problem is usually obtained by functional gradient descent. The gradient step h^t is selected in the way that $h^t \to -g^t(x,y)$, where $g^t(x,y) = \frac{\partial L(y,s)}{\partial x}|_{s=F^{t-1}(x)}$ .Often, the method utilized for the approximation is the least-squares approximation.

$$h^t = E(-g^t(x,y) - h(x))^{\,2} \tag{7}$$

Catboost adopts a decision tree as its base predictor. The decision tree divides the feature space into disjoint regions according to the values of some splitting attributes a. Splitting attributes are usually binary ones that identify that some features x^k exceeds some threshold t. This can be written as $a = I_{\{x^k > t\}}$, where x^k is either a numerical or binary feature. The final node of the tree serves as the estimate of the response y. A decision tree, therefore, can be written as:

$$h(x) = \sum_{j=1}^{J} b_j \, \mathbb{I}_{\{x \in R_j\}} \tag{8}$$

where R_j is the disjoint regions corresponding to the leaves of the tree [15].

Catboost makes use of a strategy named Ordered TS (Target Statistics) in the prevention of prediction shifts. To realize this strategy, an artificial "time", i.e., a random permutation σ of the training examples, is introduced. Then, we take $D_k = \{x_j : \sigma(j) < \sigma(k)\}$ as the training example and $D_k = D$ for a test one, where D_k is the dataset. This strategy not only uses all the training data for the learning model but also satisfies the following property:

$$E(y = v) = E(y_k = v) \tag{9}$$

where (x_k, y_k) is the k-th training example.

3.2. EVALUATION INDICATORS

After preprocessing the data, we fit the data into the models and acquired the outcome. With the purpose of evaluating the models, we picked several statistical indicators.

The first indicator is the Root Mean Square Error (RMSE). It can be utilized to measure the precision of a regression model. The way that RMSE is calculated is written as:

$$RMSE(X,h) = \sqrt{\frac{1}{m}\sum_{i=1}^{m}((h(X^{(i)}) - y^{(i)})^2} \tag{10}$$

where m is the number of instances in the dataset, $X^{(i)}$ is a vector of all feature values of the ith instance, $y^{(i)}$ is the target value for each instance, X is a matrix containing all feature values and h is the system's prediction function.

The second indicator that we picked is R-squared. The greater the value of R-squared, the better the model fits. The maximum value of R-squared is 1. R-squared is calculated in the following way:

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \tag{11}$$

where R^2 is the value of R-squared, y_i is the true value of target observation, $\hat{y}_i$ is the predicted value, and $\bar{y}$ is the mean value for the target vector.

Third, we selected the value of adjusted R-squared(adjusted R^2) because there is a problem with R^2: when the total number of features increases, the R^2 will also increase, regardless of whether the variable is indeed closely related to the target variable. Adjusted R^2 can be denoted as:

$$Adjusted\ R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2/(n-p-1)}{\sum_i (y_i - \bar{y})^2/(n-1)} \tag{12}$$

where p is the number of variables and n is the number of instances. What's more, we adopted K-fold cross-validation, as is demonstrated in Figure 8.

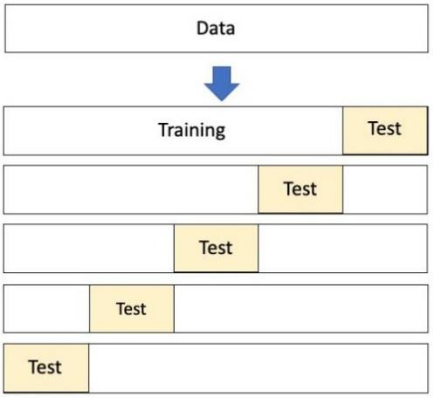


Figure 8 K-fold cross-validation.

K-fold is a vivid description of dividing the whole dataset into *K* parts with the same amount of data. During the process of K-fold cross-validation, firstly, as the name K-fold indicates, we first divide the dataset into *K* equal parts. After that we let one part serve as the test set, and the remaining parts serve as the training set. Test set is used to conduct model evaluation and see whether the model predicts well, while training set is used to train the model to fit data. Specifically, to begin with, the first part serves as the test set, and the remaining *K*-1 parts serve as the training set. Then, the second part is used as the test set, and then the third, etc. It iterates this procedure *K* times until the *K*th part has served as the test set.

4. EXPERIMENTS AND RESULTS

4.1. Model Evaluation and Selection

After determining the candidate models as well as the evaluation indicators, we processed the data with Python. During the evaluation process, R^2 and adjusted R^2 are calculated not only for the training set but also for the test set, helping us to see clearly how the models are performing respectively in the two sets. K-fold cross-validation is conducted on the whole dataset to holistically assess how well the model is performing. The results are arranged in descending order by K-fold cross-validation score, where K =5 in the experiment. Table 1 illustrates the results.

Table 1. Model Evaluation Results.

Model	Details	RMSE	R^2(training)	Adjusted R^2 (training)	R^2 (test)	Adjusted R^2 (test)	5-Fold Cross-Validation
Catboost	-	95163.23	0.954	0.954	0.912	0.911	0.91
LightGBM	-	101269.9	0.938	0.938	0.9	0.9	0.898
XGBoost	-	103746.9	0.969	0.969	0.895	0.895	0.893
Random forest	-	108767.6	0.984	0.984	0.885	0.884	0.878
Polynominal	degree=2	141416.5	0.807	0.805	0.805	0.796	0.8
Polynominal	degree=3	148391.8	0.842	0.829	0.786	0.69	0.791
Multiple	-	166881.3	0.724	0.723	0.729	0.728	0.721
Ridge	alpha=1	166877.5	0.724	0.723	0.729	0.728	0.721
Lasso	alpha=1	166881.1	0.724	0.723	0.729	0.728	0.721
Lasso	alpha=100	166871.1	0.724	0.723	0.729	0.728	0.721
Lasso	alpha=1000	167236.7	0.722	0.722	0.728	0.727	0.719
Ridge	alpha=100	167943.5	0.719	0.719	0.725	0.724	0.717
Ridge	alpha=1000	177437.5	0.685	0.685	0.693	0.692	0.683

Notes: alpha stands for regularization parameter, degree stands for the highest degree of polynomial regression, and the name in the column 'model' stands for its kind of regression, for example, Catboost stands for Catboost regression.

From Table 1, it is not difficult to conclude that Catboost Regressor performs the best among all models. It has an RMSE of 95163.23 and becomes the only model that has an RMSE of less than 100,000. When it comes to the R^2 score, adjusted R^2 score, as well as the 5-Fold Cross Validation score, Catboost stands out from the candidate models as well. Catboost demonstrates a great capability of precise prediction and does not show any tendency of overfitting, therefore, there is no doubt that Catboost is selected as the final model used to predict house prices.

The hyperparameters in the model are set by default. Here, we discuss some of the hyperparameters that are most used. In the model, the 'iteration', which means the largest number of trees, is set to be 1000. 'Learning rate' is set to be 0.03. 'Depth' means the maximum depth of the tree, which is 6. 'Class_weights' determines the

weight of each category, highly useful in hierarchical training with unbalanced data, is set to be None.

It is worth noting that the model that obtains the highest average R^2 score of the training sets in each iteration is Random Forest Regressor, which achieves an R^2 score of 0.984. However, when it comes to the average R^2 score of the training sets in each iteration, its performance is not that ideal. The average R^2 score drops to only 0.885. It is suspected reasonably that there is a slight problem of overfitting with Random Forest Regressor.

4.2. IMPORTANT FEATURES FOR DETERMINING THE HOUSE PRICE

This section will explore what features bring the most influence to the outcome of the model. The graph in Figure 9 shows feature importance generated through Catboost.

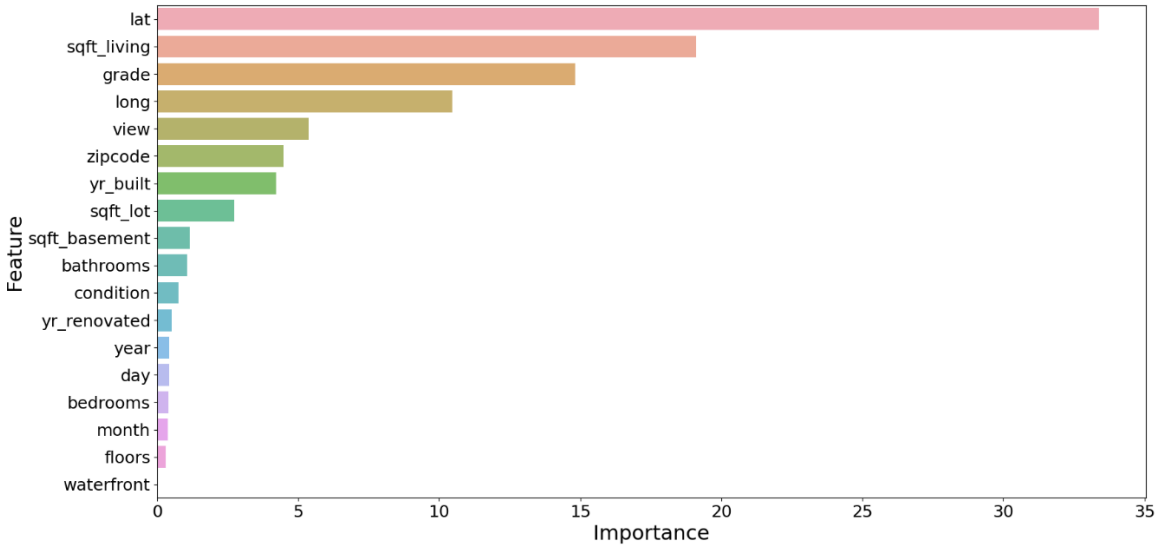


Figure 9 Feature importance graph of Catboost regressor model.

Some features that get a high score in feature importance are worth discussing. The first one is location. Note that although latitude ranks the first among all features, longitude also gets a high score and is supposed to be taken into consideration. After all, the combination of latitude and longitude represents the location of the house, thus influencing the house price. This highly conforms to real life. Location always serves as the determining factor for house price. An example for this argument prevails, for instance, a place that is convenient with public transportation is usually sold for a higher price. Likewise, a place near parks or lakes is priced highly for its surroundings. The second important factor is the area of living space. It is not surprising for 'sqft_living' to come to this place because when houses are sold, they are priced a certain amount of money per square meter or square foot. Therefore, the larger the house, the greater amount of money customers need to pay. The third important factor is the grade of the house. It is reasonable that a house in good condition will be more attractive to consumers, resulting in a higher price for sale. The feature importance outcome in this case, in general, is highly compatible with our consensus. House sellers ought to pay more attention to these features to gain more revenue and attract customers.

4.3. FURTHER DISCUSSION

This paper mainly focuses on the prediction of house prices from a scope that is comparatively micro. Properties of the houses are utilized to determine how much the houses are priced and what the important factors are in influencing house prices. A similar approach to searching for the determinant factors of house prices has also been used in a good sum of papers. A comparison and contrast of the findings of papers will be conducted in this section.

According to Mathur, who also conducted a survey of house prices in King County of the United States, the size and quality of a certain house matter the most for determining the house price. Bigger size and better quality will bring about a higher estimated value for assessors. Such a finding highly conforms to our outcome, where living space and grade rank among the top three. He also finds that a higher level of maintenance will make the house appreciate. Scholars studying other areas also contribute to this topic. Zakaria and Fatine conducted research on determinants of real estate's price in Morocco. They find out that two factors most significantly determine the house price, which is a surface area as well as the location of the real estate [16]. These two factors rank the second and the first, respectively, in our finding. Selim did research on

figuring out the determinants of house prices in Turkey. Taking even more properties into account, he concludes that the condition of the water system, whether the house has a swimming pool, and the type of the house (what material the house is made of) are the most important factors [17]. These factors seem to obviate the previous findings. However, if inspected carefully, these factors are, to some extent, related to the grade of a house. Besides, he mentions that the number of rooms and the locational characteristics is also important. These factors are compatible with our findings in this paper.

All literature mentioned above solves the problem of house price prediction and important factor determination from a microscope. Extant literature effectively attests to the validity of our paper's findings. Though there exist some slight differences, the general outcome is quite similar. House location, the space for living, as well as the condition of the house, are indeed among the most essential features from a microscope to determine how a certain accommodation will be priced.

5. CONCLUSIONS

In this paper, the issue of house price prediction is explored using a case from King County in the United States. In order to eliminate the problems that exist in the original dataset, this paper not only winsorizes the extreme values in numerical features like 'price', but also calculates the correlation coefficient and removes the highly correlated features including 'sqft_living15' and 'sqft_lot15', to assure with a precise prediction. Then, several models are utilized to fit the data. They are assessed with a variety of evaluation indicators including RMSE, R^2 score, adjusted R^2 score and cross-validation score. Among the models, Catboost outperforms all the other models and becomes the selected model because it derives the highest R^2 score and adjusted R^2 score in the test set and ranks the first in cross-validation score. The final model and corresponding essential factors are subsequently derived through Python coding. Comparison among related literature is also conducted to complete a further discussion of the topic.

From the research, we obtain the following conclusions. First, Catboost serves as the best model for our house price prediction. It not only gets the highest score in a model assessment and makes a sensible prediction, but also avoids overfitting. Second, the most important factors in the microscope that influence the house prices are location, living space and the condition of the house. Such a finding highly conforms to our common sense.

The innovations of this essay are summarized as follows. First and foremost, this essay adopts Catboost to predict house prices. This approach achieves better prediction precision compared to extant research papers on the same issue. compared to extant research papers on the same issue. In addition, this essay focuses on the house price prediction from a microscope rather than macro scope which is used by more scholars. This brings about an essential supplement to research on the house price prediction.

Despite the merits above, this essay still bears some slight drawbacks. First, this paper does not cover the macroeconomic factors. If they were taken into consideration, the results might be closer to real-life situations. Besides, this paper conducts a case study of King County of the US. However, for other areas that are not similar to King County, additional study is probably needed.

ACKNOWLEDGMENTS

REFERENCES

[1] C. Daniel, "House price fluctuations: The role of housing wealth as borrowing collateral," The Review of Economics and Statistics, vol. 95, 2021.

[2] H. Mayer, T. Sinai, "Assessing high house prices: Bubbles, fundamentals and misperceptions," The Journal of Economic Perspectives, vol. 19, pp. 67-92, 2005.

[3] H. Hirata, M. Kose, C. Otrok, M. Terrones, "Global House Price Fluctuations: Synchronization and Determinants," NBER International Seminar on Macroeconomics, vol. 9, no.1, pp. 119-166, 2013.

[4] S. Mathur, "House price impacts of construction quality and level of maintenance on a regional housing market: Evidence from King County", Housing and Society, vol. 46, no.2, pp. 57-80, 2019.

[5] T. Van, "Exploring the Advantages and Disadvantages of Machine Learning"

[6] Washington State Office of Financial Management. April 1, 2021 Population of Cities, Towns, and Counties

[7] https://www.kaggle.com/harlfoxem/housesalesprediction.

[8] R. Geiger, D. Cope, J. Ip, et al. "'Garbage in, garbage out' revisited: What do machine learning application papers report about human-labeled training data," Quantitative Science Studies, 2021; vol.2, no.3: 795–827.

[9] H. Foxwell, Creating good data: a guide to dataset structure and data representation, 1st ed. 2020.

[10] Q. Luu, M. Lau, S. Ng, and T. Chen, "Testing multiple linear regression systems with metamorphic testing," Journal of Systems and Software, vol. 182, December 2021.

[11] P. Jenny, A. Cifuentes, et al., "Towards a mathematical framework to inform neural network modelling via polynomial regression," Neural Networks, vol. 142, pp. 57-72, October 2021.

[12] L. Panzonea, A. Ulph, et al., "A ridge regression approach to estimate the relationship between landfill taxation and waste collection and disposal in England," Waste Management, vol. 129, pp. 95-110, June 2021.

[13] T. Chen, and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in the 22nd ACM SIGKDD International Conference, 2016.

[14] Q.Meng, "Lightgbm: A highly efficient gradient boosting decision tree," 2017.

[15] L. Prokhorenkova, G. Gusev, et al., "CatBoost: unbiased boosting with categorical features," 2017.

[16] F. Filali, A.Fatine, "Towards the hedonic modelling and determinants of real estates price in Morocco," Social Sciences & Humanities Open, vol. 4, no. 1, 2021.

[17] H. Selim, "Determinants of house prices in Turkey: Hedonic regression versus artificial neural network," Expert Systems with Applications, vol. 36, no. 2, Part 2, March 2021.