Feature Selection in House Price Prediction

Jia Guo *

Business College, City University of Hong Kong, Hong Kong, China

* Corresponding author: jguo44-c@my.cityu.edu.hk

Abstract. This study aims to construct a model to choose effective features and use them to predict the market price of an exact house, which can help people with house pricing and property evaluating. This paper preliminarily constructs several machine learning models like Linear Regression, SVM, KNN and compares their accuracy on this problem to choose the best-fit one for improving. After using parameter tuning to optimize this model, this study tries to use and recursive feature elimination and genetic algorithm to select features to improve simple SVM model. After feature selection, this study re-evaluated the accuracy of the model and compared which features had a greater impact on the predictions. After comparison, this study found that in this particular case, features have closer connections with living condition, traffic condition and sale condition will have a huge impact on the house price.

Keywords: Feature selection, house price prediction, SVM.

1. Introduction

1.1 Purpose

People usually predict the price of houses by collecting the features of them, such as area, garage, neighborhood and so on. While some of these characteristics can be irrelevant to or have less influence on house prices. For example, if we gather information of seasons to predict house price, the result can be unreliable even though there may be weak connection between them, so one of the purposes of this model is to pick out the most significant features and use them to train a model for prediction. Based on these predictions, people can use this result to estimate the value of properties for purposes such as mortgage loans or rental pricing. Banks can adjust loan amounts based on housing loan risk assessment and predictive model results. This can help banks accurately evaluate borrowers' repayment capacity and loan risks, enabling them to formulate more reasonable loan policies and interest rates. For individuals and families planning to buy a house, housing price predictions can provide valuable guidance. Based on the prediction results, they can make more informed purchasing decisions, such as determining the optimal buying timing or selecting a housing type and location that fits their budget.

On the other hand, this study also helps to find out what is the main content of the house price. Those features having obvious affection to prices indicate the expectation of consumers to houses. This prediction can give recommendations about which part of the houses should be highly advertised to attract buyers, because they are the most concerned about. And it is a cost-saving way to focus on these features when constructing houses.

1.2 Literature Review

Traditional methods for predicting housing prices mainly use statistical methods and economic models, such as regression analysis, time series analysis, and impulse response functions. These methods are based on historical data and economic indicators to model and predict housing prices. However, traditional methods have issues such as high data requirements, strict model assumptions, and sensitivity to changes in the economic environment.

Nowadays, machine learning getting more and more popular to predict the house price. Regression models are one of the most used methods in the field of machine learning. Common regression models include linear regression, polynomial regression, and ridge regression. These models predict future

housing prices by fitting the variable relationships in historical data. However, regression models have certain limitations when dealing with high-dimensional data and nonlinear relationships.

There have been many previous studies on housing price prediction, in previous study, researchers tried to analyze house price from different aspects. In Zhu's study, they constructed a relationship between house price with POI statistics to study this problem in a certain region [1]. And Yang tried to use development of railway system of a city to explain the change of house price in that city [2]. In addition, in the study of Li, they mainly focus on the effect of macroeconomy and administrative factors [3]. While there is less study on using the features of houses themselves to predict price before. The information of a certain house is various thus the features are complicated.

When we deal with multidimensional features, some advanced algorithm like decision trees and SVM are commonly used. However, some of these models are prone to overfitting, thus require appropriate parameter adjustments and model optimization.

So, this project aims to use feature-selecting and Parameter adjustment to Improve the generalization ability of the model.

2. Methodology

2.1 Data Processing

Data of house price and influence factors are obtained from the website and are imported as pandas Dataframe for further handling. Next step is to perform a basic processing on the data, after a brief examination of the data, it is discovered that there are a significant number of missing values in the dataset, and these missing values mainly appear in several features, so it is better to delete those columns having too much missing values. In addition, for those features having fewer missing values, this article uses average value to replace missing value.

After that, the dataset has been transformed into a new Data frame which contains 62 features with 34 numerical features and 28 categorical features.

The heatmap can clearly display which features have significate influence on sale price. the closer to 1 absolute value of covariance between features and sale price is, the bigger influence features have on sale price. This gives a suggestion for primary feature selecting to construct models (See Fig. 1).

After observation of the dataset, there appears a problem that some categorical features have too many kinds of objects or just have one kind of objects. So, this study only chooses those categorical features with less than 4 and more than 1 kind of objects as important columns. To preliminary construct models, this article first sets 0.5 as the boundary condition to select features. Characteristics whose absolute covariance between sale price is more than 0.5 are chosen to form a new data frame with important categorical features. To switch the boundary condition to select different features for model training and can find out how different features selected will influence price prediction.

To improve the accuracy of models, it is also necessary to standardize the data to decrease the influence of outliers in the data on the model and scale of data.

One-hot encoding makes it possible to transform categorical data into numerical for model training. It is necessary to turn each option under every feature into a new feature, and then use 0 and 1 to represent whether this option is chosen or not. This provides a method to handle with categorical data, although this will elevate the dimension, the value under each column will be easier to deal with.

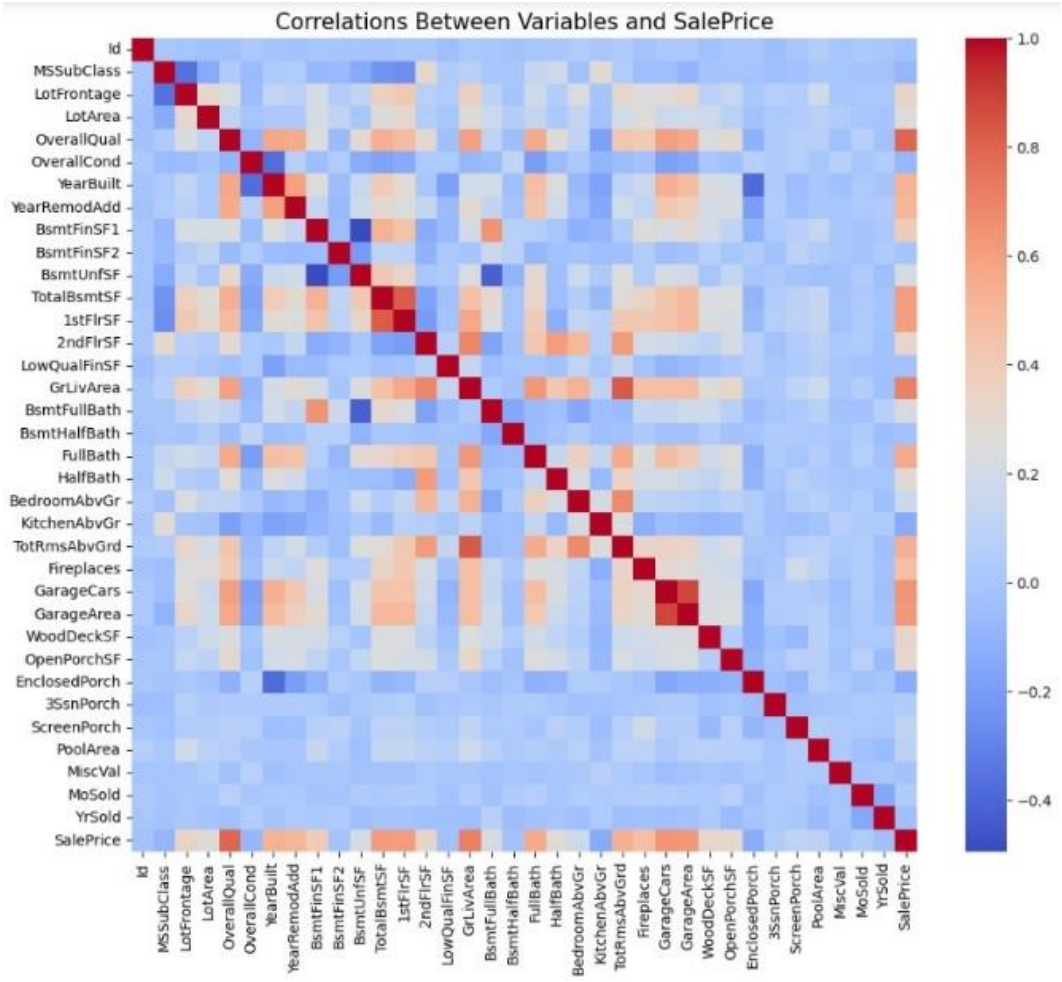

Figure 1. Correlations between variables and Sale Price

2.2 Model Training

At first step, several different models are separately trained to predict the house price. Then the result of prediction will be evaluated and measured on the standard of MSE. In this study, Linear Regression, Elastic Net, Random Forest, SVM, XGBoost, KNN model are chosen for the Preliminary test.

The Elastic Net algorithm is able to handle situations where there are a mass of features or predictors compared to the number of observations in the dataset. This algorithm can be a combination of two widely used regularization methods: L1 regularization (Lasso) and L2 regularization (Ridge). L1 regularization encourages sparsity by shrinking some values to exactly zero, which can act as the part to select feature. While L2 regularization will have other function to reduce the impact of multicollinearity by shrinking the values towards zero. It can automatically select relevant features and reduce the model's reliance on irrelevant or redundant predictors.

One reason to use Decision tree model for house price predicting is probably benefit from its advantages in dealing with non-linear relationships and interactions between features. Decision tree is a non-parametric model, which means it only raises very few hypotheses about the underlying data distribution. This flexibility allows decision trees to handle complex relationships that might exist between the predictors and the target variable [4]. Based on bagging, further some randomness is additionally introduced in the training process of decision trees: when it comes to divide attributes, the system will first randomly select a subset containing k attributes, before selecting the optimal one from the subset, the attributes are divided, which is also the connotation of randomness. But this sacrifices interpretability at the same time. Gradient boosted trees have a component of multiple decision trees as well, but what makes it different is that where trees are trained in parallel on

bootstrap samples of the original data set, the trees are trained in a sequential manner [5]. The optimization criterion of XGBoost is mainly coming from the minimization of the objective function, which incorporates regular terms in each iteration to reduce overfitting risks. Unlike the heuristic iterative principle of GBDT, XGBoost employs a second-order Taylor expansion and allows for customization of the loss function. This further enhances its performance compared to GBDT.

SVM was originally used for classification, but its principle was extended to the task of regression and forecast as well [6]. Support Vector Machines (SVM) has advantages on dealing with high-dimensional data and capture complex relationships between features and the target variable. SVM is especially effective in cases where the number of features is large but the samples is not large enough to for precise model training. In order to make it easier to construct a hyperplane which can classify all data, there is a method named kernel trick is used to place data in higher-dimensional space.This can be beneficial when dealing with datasets that have many predictors. Moreover, SVM are known for their robustness to outliers in the data. This algorithm aims to find the condition that exactly make the rate between different classes the maximum, which reduces the influence of individual data points. This can be advantageous in-house price prediction, as outliers or extreme values in the predictors may not significantly affect the model's performance.

KNN is a non-parametric algorithm and one of its features is that it does not make any strong hypothesis about the basic data distribution. It works by finding the k closest data points in the training set to a given test data point and making predictions based on the majority class or average of their target values. Compared with other nonlinear models for house price prediction, one of the greatest advantages of neural networks is that a class of multilayer neural networks could approximate a large class of functions well [7]. Another advantage of KNN is its ability to adapt to local patterns in the data. Since the prediction is based on the nearest neighbors, the algorithm can capture localized trends and variations in house prices. This can be valuable in scenarios where house prices may vary significantly across different neighborhoods or areas.

This study uses these algorithms from scikit-learn library to train the models. After that, in order to compare the accuracy of different models, this article uses evaluate them by calculating RMSE in cross validation (See Table 1).

Table 1. Evaluation of Prediction of Different Models

Model	MAE	MSE	RMSE	R2 Score	RMSE(Cross-validation)
SVM	16369.730983	562670700	23720.681411	0.895832	29500.588856
XGBoost	18489.425367	745131500	27297.096605	0.862053	29583.291523
RandomForest	18060.898116	704352200	26539.633986	0.869603	32094.596835
LinearRegression	21040.608456	817825400	28597.646151	0.848595	35946.489817
ElasticNet	21283.207214	863888900	29391.986160	0.840068	38463.841711
KNN	27543.568493	1796712000	42387.640920	0.667373	38523.843496

In the preliminary test, the SVM performs better than other algorithms. This might because that using SVM as the underlying algorithm allows for robust and accurate predictions by effectively capturing complex relationships in the data. The flexibility of SVM enables it to handle non-linear patterns and high-dimensional feature spaces commonly found in house price datasets. And since my dataset is not large enough, SVM can have a more satisfying result when handling small samples.

2.3 Hyper-parameter Optimization

Before selecting features, to further promote the accuracy of model, this study chose the method of random search to readjust the parameter. Random Search will be more efficient than Grid Search for hyper-parameter optimization. According to Bergstra and Bengio, compared to other neural network models applied with pure grid search, random search can find equally good or better models in a more time-saving way, when the domain of parameters is the same. In case that the computational budget is limited, random search will explore a larger, less promising configuration space effectively on purpose to find better models [8]. By randomly sampling hyperparameters, Random Search has a

higher chance of finding optimal or near-optimal combinations, especially when the search space is large and complex. This can be advantageous in-house price prediction, where different models or algorithms may have multiple hyperparameters that need to be adjusted for optimal performance. Furthermore, Random Search is easy to implement and does not require prior knowledge or assumptions about the data. It has a significant versatility that can be widely combined with different models and algorithms for house price prediction.

2.4 Feature Selecting

In the previous model training, the selection of features still needs to be optimized. In previous training, all the features are filtered based on covariance. When adjusting the boundary condition from 0.4 to 0.7, the score of models also changed. Feature selecting does have significant influence on prediction. And the model performs the best with choosing 0.4 as the condition when there are 10 features are selected. As a result, at most 10 features will be selected as well in further study. While simply choosing features by comparing the covariance is rough. So, in the next step, this study will use genetic algorithm (GA) and recursive feature elimination (RFE) to select features for model training (See Table 2).

Table 2. Result in Different Conditions of Covariance

Covariance	MAE	MSE	RMSE	R2 Score	RMSE(Cross-validation)
0.4	16346.699427	550915800	23471.595720	0.898009	28906.453160
0.5	16369.730983	562670700	23720.681411	0.895832	29500.588856
0.6	17169.348623	571478200	23905.609569	0.894202	29788.248143
0.7	21602.283228	838995300	28965.414829	0.844676	34194.046495

The optimal configuration, given a set of features, is the collection or subset of those features. This approach involves discrete selection. The cost of determining the optimal feature set can be very high when dealing with a large number of possible permutations. Genetic algorithm employs an evolutionary-based method to determine the optimal set. When using GA for the purpose of feature selection, it has to construct a population of subsets which concluding all the features in dataset at first. And then each subset in the population will be evaluated by verifying its actual effect in the chosen prediction model completing the target. Once the members of the population are assessed, a competition is held to determine which subsets will continue to the next generation. The next generation is composed of the winners from the competition and undergoes crossover (updating the winning feature set with features from other winners) and mutation (randomly introducing or removing some features). The main advantages of genetic algorithms for feature selection include strong robustness, simplicity, and generality, according to Gao [9]. The most remarkable advantage of GA can be that it performs well in dealing with complex problems and can be applied to parallel and distributed processing.

RFE also belongs to wrapper-type feature selection method in which the primary principle is that using machine learning model to score each subset of features and selecting the better ones. Each feature's importance score is measured by how the prediction fits the real value, and among all the features, the one with the lowest is deleted from the subset for it may have bigger misleading to the model. This process is repeated until the number of features is reduced to the preset number [10]. Recursive Feature Elimination helps to identify the most informative features by iteratively evaluating their impact on the model's performance. By considering the feature importance within the context of the model, RFE can have a good effect on understanding complex relationships and interactions between features that may not be apparent in a standalone analysis.

3. Result

These 2 methods pick up 10 features to predict the house price, and the 10 features is the same with those chosen basing on the criterion of covariance. To verify the influence of new chosen

features and to test the new parameters' score, this study trained the model again with new statistics. The result is as shown in Fig. 2, in most situations, it fits well. Except for several special cases, it gives a satisfying prediction. Besides, it is noticeable that those values which differ a lot from predictions are mostly located at the ends of the price range. So, the error might be due to the outlier in the dataset (See Fig. 2 and Table 3).

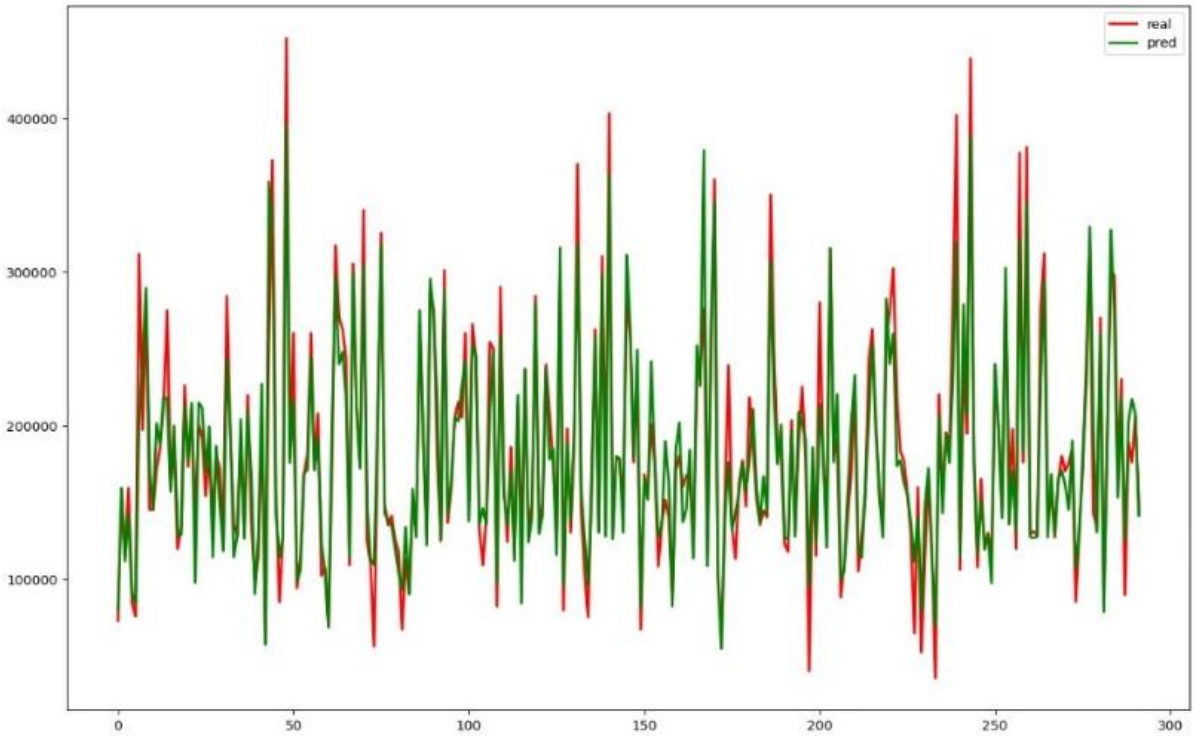

Figure 2. Prediction Result

Table 3. Result after optimization

	MAE	MSE	RMSE	R2 Score	RMSE(Cross-validation)
Original	16369.730983	562670700	23720.681411	0.895832	29500.588856
Improved	14573.412343	426424800	20650.055689	0.924431	26384.734457

The new model is also evaluated and compared with the original one. Using these features to train the model again with adjusted parameters, RMSE gets smaller, the prediction is more accurate. The optimization does improve the SVM model.

In addition, this experiment reveals that the most influential 10 features are: total area, garage capacity, construction time, garage area, heating, kitchen area, utilities, residential type, land slope and sale condition. RFE operates within the feature space and aims to find the best subset of features for a specific model. While GA explores a larger search space that includes different combinations of features and can potentially find global optima. However, in simpler problem like in this case, although both of them give the same result but GA is more computationally intensive for involving operations like selection, crossover, and mutation.

4. Conclusion and Future Work

4.1 Conclusion

This study tried to use different model for house price prediction. After comparation with Linear Regression, Elastic Net, Random Forest, XGBoost, SVM and KNN, SVM is found to be the most suitable model in this case and is chosen for further optimization. This study then uses SVM model combined with RFE and GA for house price forecasting. It aims to just use the original properties of

the house itself to predict the price. After feature selection, it is found that total area, garage capacity, construction time, garage area, heating, kitchen area, utilities, residential type, land slope, sale condition make a difference in house price. These important characteristics can be concluded to three categories: living condition, traffic condition and sale condition. Although the appearance of the houses may also be concerned by house buyers, it turned out to be not that important to prices.

4.2 Future Study

In the real life, when gathering the features of houses, people may not be able to obtain all needed information of a house, in that case my model will not work. People have to construct a new model to predict which takes time. In the future study I aims to conduct a model which can adjust needed conditions automatically depending on users' needs. It will perform better when some features are missing.

At present, there are many methods for house price prediction, including regression analysis, artificial intelligence, machine learning and other methods. However, the accuracy and stability of these methods need to be further improved. In addition, house price prediction is also affected by many factors such as economic policies, financial market, real estate market and other factors. Anyway, it is difficult to predict the future house price with a single method.

References

[1] Zhu Haiyu, Wang Hijie, Ye. Cancan Prediction of housing prices in urban hotspot areas based on XGBoost algorithm - taking Nanjing Jiangbei New District as an example. Construction Economics, 2022, (S2): 433 - 437.

[2] Yang Linchuan, Chen Yang, Xu Nenglai, Zhao Rui, Chau K.W., Hong Shijian. Place-varying impacts of urban rail transit on property prices in Shenzhen, China: Insights for value capture, Sustainable Cities and Society, 2020, 58: 102140.

[3] Li, N., Li, R. Y. M., & Nuttapong, J. Factors affect the housing prices in China: a systematic review of papers indexed in Chinese Science Citation Database. Property Management, 2022, 40 (5): 780 – 796.

[4] Zhou, Z. H. Machine learning. Springer, 2021.

[5] Anders Hjort, Johan Pensar, Ida Scheel & Dag Einar Sommervoll. House price prediction with gradient boosted trees under different loss functions, Journal of Property Research, 2022, 39 (4): 338 – 364.

[6] Wang Xibin, Wen Junhao, Zhang Yihao, Wang Yubiao Real estate price forecasting based on SVM optimized by PSO. Optik, 2014, 125 (3): 1439 - 1443.

[7] Wang Tao, Yang Jian. Nonlinearity and intraday efficiency tests on energy futures markets, Energy Economics, 2010, 32 (2): 496 - 503.

[8] Bergstra, J., Bengio, Y. Random search for hyper-parameter optimization. Journal of machine learning research, 2012, 13 (2).

[9] Gao, Y. M., Zhang, R. J. Analysis of House Price Prediction Based on Genetic Algorithm and BP Neural Network. Computer Engineering, 2014, 40 (4): 187 – 191.

[10] Lamba, R., Gulati, T., & Jain, A. A Hybrid Feature Selection Approach for Parkinson's Detection Based on Mutual Information Gain and Recursive Feature Elimination. Arabian Journal for Science and Engineering, 2022, 47 (8): 10263 – 10276.