Integrating Machine Learning Techniques for Real Estate Analysis

Zekun Chen *

School of computing and augmenting intelligence, Arizona State University, Arizona State, USA

* Corresponding Author Email: zchen409@asu.edu

Abstract. Machine learning can be applied to prediction models. The study delved into house price prediction, an area of significance to both individual participants and policymakers within and beyond the real estate market. The advanced predictive model is a tool that informs quality decision-making. This study synthesizes and builds on existing research to implement and enhance house price prediction methods, assessing the performance of Convolutional Neural Networks (CNNs), Decision Trees, and K-Nearest Neighbors (KNN) which is of focus on Most Correlated Features (KNN-MCF). Although these techniques have been developed and refined over the years, they still face challenges such as overfitting, noises existing in the dataset, and the complexity of modeling both linear and non-linear relationships. By integrating the resources, the study aims to provide objective guides for addressing the common challenges and predicting house prices, illustrating the appropriate analytical methods to apply based on characteristics of a database, extra space for improvements, and critical concerns of their practical applications. Ultimately, the research strives to bridge the gap between theoretical models and real-world applications to develop practical tools for reliable house price prediction.

Keywords: KNN, CNN, Decision Tree, House price prediction, Application.

1. Introduction

The real estate market is known not only as a fundamental necessity, but a common investment for wealth creation and preservation. In the real estate market, both individual subjects like buyer and seller and government are involved. As the most basic and important information in such a market, the house price is expected to be predicted as assistance for all subjects involved to make informed and prudent decisions. To achieve the best reasonability for decision-making, all subjects involved in the market expect scientific methods to predict house prices because the scientific method reduces bias by relying on objective evidence to keep accuracy. Informed with the information about predicted house prices, buyers and investors will be aided with risk management, and the government will be able to allocate the resources more wisely such as funds for infrastructure and neighborhood development. The scientific prediction will assist that planning and strategies be grounded with evidence to achieve a higher likelihood of intended outcomes.

Several models and techniques are popularly used in housing price prediction, depending on machine learning and deep learning. Most of the existing models that are used to predict the house price is seeking for the relationship between provided data and house price. Convolutional Neural Networks (CNNs) are widely used in processing visual data to extract features for prediction applications. Additionally, the more structured house price data that CNNs transform from imagery can be used for further analysis and prediction. Decision Tree is an approach that is flexible on data. Practically, the provided data is often a mix of categorical and numerical ones. The function of a tree structure that can process both two kinds of data at the same time reduces the difficulty of data processing, which is a helpful tool in handling given heterogeneous housing data and concluding the higher-level relationship between data. As an advanced tool of decision trees, Random Forest constructs a multitude of decision trees every training time, which provides higher accuracy and reduces the risks of overfitting. K-Nearest Neighbors (KNN) is famous for its simplicity and intuitiveness, with the notable advantage that no prior training phase is needed so that the model can

remain unbiased by previous data. With no prior training phase, KNN is particularly helpful when the data it involves is abundant but continually evolves or fluctuate frequently.

This study analyzes the application of the abovementioned methodologies in other studies, discussing their structures and functions in the context of house price prediction. While assessing the performance of each methodology, the study shows particular importance on data handling and feature selection in preprocessing. From the analysis, the comparison is made among models so that we learn that different models excel in different contexts. When practical application challenges exist for consideration, tips for improving the house price prediction model can be adopted to achieve better prediction results, contributing to more prudent decision-making.

2. Overview of machine learning techniques: CNNs, Decision Trees, and KNNs

Accurate house price prediction requires powerful tools. Among the renowned machine learning algorithms, CNNs, Decision Trees, and KNNs are particularly popular for their effectiveness and accuracy.

CNNs typically include convolutional layers, pooling layers, and connected layers. Convolutional layers apply a set of filters to the input image to create a feature map that represents the presence of specific features within the input. The pooling layer follows the convolutional layers and aims at reducing the spatial size of the representation and extracting higher-level features from the input. Combining the filtered features gained from the previous layers to form a model, a fully connected layer applies a series of weighted connections and activation functions that enable the network to make complex decisions and predictions. When predicting the house price, convolutional layers help in identifying visual features from images of houses that might be indicative of prices, like the size of the house. Pooling layers decreases the complexity of computation and the number of features to process, when the minor differences in visual features will be of less significance in the prediction work (like the variation of colors), allowing the model to focus on the characteristics of the house of more importance for valuation (like the presence of a swimming pool). The fully connected layer is where actual regression prediction takes place. They integrate the features extracted from previous layers into the form that the network can use, compute the weighted sum of its inputs determine the neurons' outputs, and adjust the weights to learn which feature is most indicative of the house price [1]. For instance, the weights can determine how much the size of a house contributes to its price so that the network can adjust the weights for price prediction accordingly. Activation layer enables the network to learn complex features from the input.

A decision tree is a method of recursive partition. Recursively, the algorithm first selects the attributes on specific criteria that separate the samples to distinct classes from the training data, and then create subsets according to the attributes. The recursive partitioning ends when stopping conditions are met, which marks the point when the tree has been constructed. The decision making for new data is carried out when the new data follows the decision rules learned during the tree's construction to a leaf node that represents the outcome of the input data [2]. Practically in the context of house price prediction, the algorithm would select attributes from the dataset that are most indicative of house prices, such as the location, age of property, and number of bedrooms. Recursively, in the subset created according to the criteria of the specific location, the algorithm can split the subset further according to the age of the property so that a more homogeneous subset of houses will be found. The recursion partition continues util meeting a certain condition, such as the subsets cannot be split any further in a meaningful way (like when subsets are of very similar features) or the expectation for the tree has been achieved. When predicting the house price based on new data, the developed tree is able to navigate the data from root to a leaf node. The arrival at leaf node is when the prediction of house is completed.

K-Nearest Neighbors (KNN) can be applied for both classification and regression tasks, predicting the new cases based on how closely the new cases resemble the existing cases. "K" in KNN represents selecting the number of neighbors that would be grouped together with the new data for analysis.

When “K” is decided, the next step is typically calculating the distance from the new data point to every other point in the training dataset. Based on the selected number of neighbors and the ranking of distance, the nearest neighbors can be selected for prediction use, which involves the methods such as majority vote in the classification task and averaging in the regression task. Since the house price data are usually continuous numerical values, the prediction of house prices typically falls under regression task. Choosing the value of “K” is a crucial step as it thoroughly impacts the performance of the model. Good strategy is combining Cross-Validation and the Square Root Rule: when starting with assigning to “K” the result value of the square root of entire data size, the algorithm splits the dataset into training sets and validation sets so that the model can be trained with different values of “K” to find the optimal “K” for the best performance (Fig.1). The combination of Cross-Validation and the Square Root ensures the balance that model performance is evaluated across sufficient subset cases and the values of “K”, which helps that overfitting is effectively avoided. Once the value of “K” is determined, distance calculation, nearest neighbors’ selection, and regression prediction should follow to achieve the house price prediction. Regression prediction predicts a value usually by averaging or taking the median of the output values of nearest neighbors of new data.

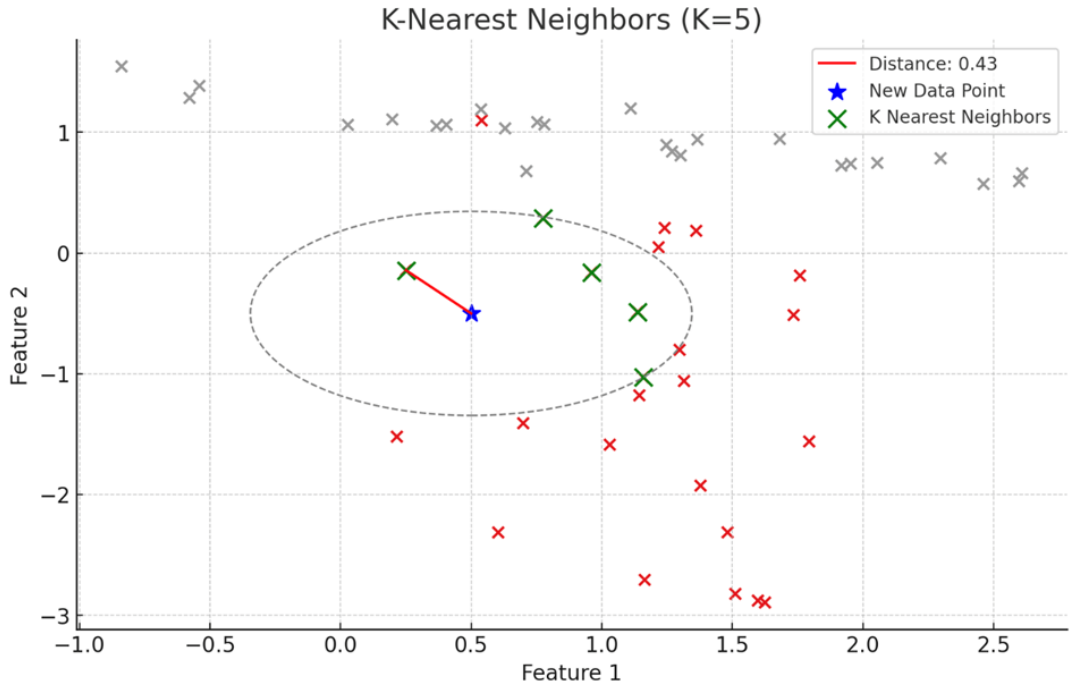

Figure 1. Description of KNNs

3. Application of Modeling Techniques in House Price Prediction

Yong Piao; Ansheng Chen; Zhengdong Shang studied on a predictive model tailored for housing prices employing a Convolutional Neural Network (CNN) and using the dataset of real estate transactions from Dalian, China [3]. The study starts with preparing and cleaning the extensive dataset, and adopts XGBoost Model to determine the most influential attributes for feature selection. The CNN model is structured with two convolutional layer and includes dropout methods to avoid overfitting. Designed for continuous value regression, this model demonstrates high accuracy. The dataset adopted in this study is large and complex. XGBoost featured with its optimized computing power to handle large dataset as well as its ensemble learning approach to uncover complex patterns of relationship between variables justify its suitability for feature selection in this case. Additionally, XGBoost can scale seamlessly with increasing data, which is helpful to maintain its accuracy and performance when more data of real estate would be available. Two Convolutional Layers provide balance between simplicity and complexity of feature recognition, which maximizes feature extraction while keeping computational demands manageable. The strategy adopted to avoid

overfitting is incorporating dropout layer through randomly deactivating a subset of neurons in a neural network during training to ensure the prediction models work with generalization ability to new data other than training data. Typically, CNN are widely used for image recognition and classification tasks, but to achieve the numerical tasks for numerical outputs like house price, CNN is adapted for continuous value regression. With the process of hierarchical learning, CNNs can understand the inputs at various levels: from basic attributes like size and location to more complex interactions such as the combined effect of neighborhood and economic trends. The extraction of CNNs can identify intricate patterns and relationships among input attributes which is hard to discern manually so that more comprehensive analysis can be considered for predicting results. However, to keep accuracy of more general and boarder attributes is more challenging. Practically, choosing the right kernel size and stride is crucial: while larger kernels and strides are beneficial for larger market trends. Smaller kernels and strides are beneficial for detailed information like individual properties.

Zhishuo Zhang in his paper developed a model to predict the house price employing a decision tree regressor [4]. The researcher selected important features from the Boston housing dataset, applied the Classification and Regression Tree (CART) algorithm, and adopted rigid search and cross-validation to optimize the decision tree parameters. The achieved predictive model turns out to be successfully of high accuracy. Housing prices are influenced by a multitude of interrelated factors, and decision trees can effectively process these non-linear interactions without complex preprocessing and feature engineering. Selecting the most informative feature from the dataset helps the model to focus on the most relevant predictors and ignore the noise. Zhang's research is designed to select the most critical features from the Boston housing dataset, such as the number of rooms, lower status of the population, per capita crime rate, geographic location, and student-researcher ratio, which are identified as the most impactful features to the house price. CART exclusively uses binary splits (dividing the data into two groups at each node), which simplifies the tree structure and contributes to its flexibility of handling mixed categorical and continuous variables efficiently from Boston housing dataset. By exploring a grid of parameter combinations and optimizing through cross-validation, the model effectively avoids overfitting, enhancing the prediction accuracy. With binary splits, pruning mechanisms, no need for data preprocessing, CART requires less computation power and simpler to implement when compared with other models or algorithms. However, although good at non-linear relationships, the decision trees may not be the optimal method in analyzing the datasets of linear relationships when other models like linear regression is available. Other shortcomings of decision trees include the limited capability of extrapolating to predict house prices in scenarios largely different than the training data, and its vulnerability to small changes in the training data which may incur different paths and different prediction results.

In their paper, Karshiev Sanjar and colleagues introduced a novel application of K-Nearest Neighbors algorithm, refined to enhance house price prediction accuracy [5]. The technique employed is called KNN based on Most Correlated Features (KNN-MCF). MCF focuses on only the most influential factors affecting house prices to mitigate the negative impact of missing information on KNN's performance. With benefits of MCF, KNN identifies the "k" closest neighbors based on the known features that are most correlated to the missing data and uses those neighbors to estimate the missing values in the dataset. Once the missing data has been imputed, the resulting refined dataset can be used to train various machine learning models to predict house prices for more accurate predictions. In their paper, the consequence of missing data reduces the sample size, compromises the statistical power of analyses, and potentially leads to biased estimations, which reveals the needs that the missing values be processed appropriately. In their paper, more accurate imputation for the missing data can be generated with KNN when adopting the values from similar instances from existing dataset, rather than a general metric like the mean or median. Usually, house prices are significantly influenced by local trends and other more general characteristics in the estate markets, which can be captured and concluded when KNN looks at the nearest properties of dataset, achieving more comprehensive evaluations for prediction. In addition to its role in preprocessing the datasets, KNN can be a standalone model for predicting house prices through identifying the nearest neighbors

of the target property, provided that the data has been well structured. While demonstrating high accuracy, the application of KNN-MCF should be considered with its risks. KNN requires intensive computation and performs a phenomenon known as the curse of "dimensionality" when dealing with large dataset. While MCF typically is limited to linear correlation, its capability is limited for problems which are more complex and of non-linear relationship. Moreover, the dataset processed by MCF may oversimplify the dynamics and interaction effects between data. For example, the dynamic between "proximity to the city center" and "number of bedrooms" may interact in a way that their combined effect on house price is greater than the sum of their individual effects.

4. Conclusion

The studies on implementing multiple models provide us with insights about how to achieve better performance of the model. The development of a predictive model for house pricing requires meticulous attention to each phase of the modeling process. This involves understanding the characteristics of each model and database, the strategic selection of features, and the implementation of specific methodologies for each model. To optimize the performance of prediction outcomes, one must carefully curate datasets that are both comprehensive and representative of larger trends. This process includes gathering the raw factors affecting the price from different aspects and minimizing the impact of outliers and exceptions on prediction results. Mitigating underfitting, overfitting, and bias is necessary to enhance model reliability through rigorous cross-validation techniques and regularization methods. By seeking the interrelations among housing attributes, one can establish more refined model rules to improve the outcome accuracy. The choice of model should match the complexity and structure of the database, with appropriate performance metrics selected for further adjustments to the models.

Ensemble models can significantly enhance the accuracy and stability of price prediction by pooling unique advantages of different algorithms, such as the feature detection capability of CNNs and the imputation capability of KNNs. Their work together effectively averages out the biases and variances of individual models, thereby reducing the likelihood of underfitting and overfitting, which offer predictions better reflecting real-world trends. Collecting the insights of various models, ensemble models can achieve superior predictive performance, metaphorically seen as arriving at a consensus. However, the trade-off between complexity and performance should be taken into consideration when deciding the application of ensemble models as more complex models can yield more accurate results while it is more challenging to implement.

References

[1] Wu, Jianxin.Introduction to convolutional neural networks. National Key Lab for Novel Software Technology. Nanjing University. China, 2017, 5, 23: 495.

[2] De Ville, Barry. Decision trees. Wiley Interdisciplinary Reviews: Computational Statistics 2013, 5 (6): 448 - 455.

[3] Piao, Yong, Ansheng Chen, and Zhendong Shang. Housing price prediction based on CNN. 2019 9th international conference on information science and technology (ICIST). IEEE, 2019.

[4] Zhang, Zhishuo. Decision Trees for Objective House Price Prediction. 2021 3rd International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). IEEE, 2021.

[5] Sanjar, Karshiev, et al. Missing data imputation for geolocation-based price prediction using KNN–MCF method." ISPRS International Journal of Geo-Information 9.4 (2020): 227Fangfang. Research on power load forecasting based on Improved BP neural network. Harbin Institute of Technology, 2011.