Machine Learning for Housing Price Prediction

Rahimberdi Annamoradnejad
University of Mazandaran, Iran

Issa Annamoradnejad
Sharif University of Technology, Iran

INTRODUCTION

The housing market is one of the earliest and most influential industries with interests among general populations. It has been described as the least transparent industry in our ecosystem, as it keeps changing day in and day out (Varma et al., 2018). In recent years and with the advent of computer approaches, many studies used the latest machine learning models to analyze the housing market and identify its most important influential variables in order to suggest a proper price or to predict price fluctuations. These automated models help homeowners, builders, and business people to perform an objective assessment and reach more profit by knowing about the future market, successful neighborhoods, or demanded building structures. In addition, they assist city administrators and urban planners to organize environmental and locational amenities in a way that would bring success and profit to a wider population of a city.

Earlier works were able to apply statistical and mathematical approaches, such as regression and hedonic price models (Harrison Jr. & Rubinfeld, 1978), on relatively small datasets collected from local real estate agencies. Throughout the years, precision was significantly improved by collecting larger datasets and variables, and with the use of the latest machine learning models, e.g. neural networks, deep learning, and gradient boosting. Most previous studies consider property price prediction as a static task, without any regard for price fluctuations over time. However, in a real-world setting, it is also essential to include time in making a final prediction. Stock market prediction is a similar but more volatile example, where researchers focus on predicting price changes based on daily data.

This chapter follows the general phases of the CRISP-DM process model for data mining, to elaborate on the problem statements, data collection and preparation, modeling, and evaluation. It will take into account the intricacies of the problem in all steps of the process, from the problem definition, data collection, feature engineering, model selection, and evaluation. The discussion contains a classification of tasks, popular machine learning competitions, variable groups, and feature engineering methods. Data sources and methods of data collection will be reviewed in detail and variables are grouped into three categories. Proper ways to design steady and accurate models are proposed in relation to previous methods and approaches for predicting housing prices. While the purpose of this chapter is not to propose a new model or dataset for housing prices, the overall workflow and proposed ideas can empower future studies to design a machine learning approach for predicting housing prices.

DOI: 10.4018/978-1-7998-9220-5.ch163

BACKGROUND

Early works studied housing prices using mathematical and statistical approaches. The hedonic pricing is a very popular method in early works of determining influential variables of housing prices, which detects the impact of given variables on the total price of a property and is used for valuation of market goods for their utility-bearing characteristics (Harrison Jr. & Rubinfeld, 1978; Selim, 2011). This conventional method has been widely applied for housing prices (Tse, 2002; Hansen, 2009; Selim, 2011; Liao and Wang, 2012).

In recent decades and with the advent of computer models, many studies applied state of the art models to analyze the housing market, identify its most influential variables, and suggest a proper price. Many machine learning algorithms and approaches have been applied to assess their success in making accurate predictions. Earlier works focused on training general-purpose statistical and machine learning models, such as Fuzzy logic (Kusan et al., 2010), Neural Networks (Nghiep, 2001; Khalafallah, 2008), Linear Regression (Sangani et al, 2017), Random Forest (Antipov and Pokryshevskaya, 2012), Support Vector Machines (Kontrimas and Verikas, 2011), Genetic algorithms (Giudice et al., 2017) and their blended results (Kontrimas and Verikas, 2011; Plakandaras et al., 2015). Antipov and Pokryshevskaya (2012) addressed the missing values problem in most housing variables datasets and showed that the random forest approach is suitable to deal with them. In recent years and with the possibility of creating large datasets, works utilized more complex neural networks (Pai and Wang, 2020), deep learning, and gradient boosting (Sangani et al., 2017; Singh and Sharma, 2020) to tackle the problem. Truong et al. (2020) adopted Stacked Generalization approach, among other ensemble techniques, to optimize the predicted values.

Most previous studies and machine learning competitions consider housing price prediction as a static task, without any consideration for price fluctuations over time. While that is a valuable goal to understand the essential variables on the final pricing, it is also important to predict the overall price changes and consider market fluctuations to reduce external bias from the training. Stock market prediction is a similar but more volatile example, in which many researchers have tried to predict price changes in day to day markets. To this aim, some works applied time series methods, autoregressive and dynamic models to forecast house price growth rates and volatility (Segnon et al., 2020; Bork and Møller, 2015). Plakandaras et al. (2015) designed a model to detect early warning signs for predicting sudden house price drops. Previous studies are mostly fine-tuned on a single dataset of a particular city, as with the several machine learning competitions that target predicting housing prices.

PREDICTING PRICE VALUES

This section discusses the necessary steps in developing a new machine learning model for housing prices, according to CRISP-DM methodology (Wirth & Hipp, 2000) and in line with the general steps of supervised learning algorithms (as described by Kotsiantis et al., 2007). The CRISP-DM process model is shown at Figure 1. The discussion contains exploration of problem definitions, dataset creation methodologies, variable groups, proper feature engineering methods, modeling, and evaluation. We omit the last phase of CRISM-DM, i.e. deployment, in this chapter.