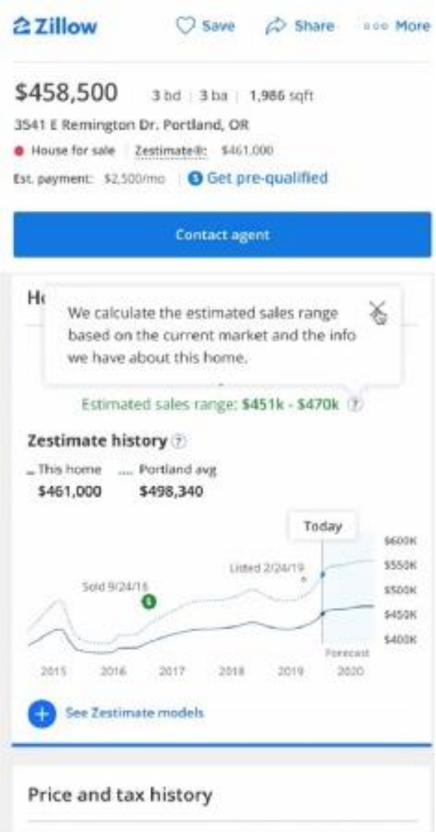


$458,500 3 bd | 3 ba | 1,986 sqft

3541 E Remington Dr. Portland, OR

● House for sale | Zestimate®: $461,000

Est. payment: $2,500/mo | Get pre-qualified

We calculate the estimated sales range based on the current market and the info we have about this home.

Estimated sales range: $451k - $470k

Zestimate history

This home	Portland avg
$461,000	$498,340

Price and tax history

Data Mining and Machine Learning in the context of Estimating Housing Costs

An Informative Description to Help New Home Buyers

Lukas Nilsen | ENL 3030

Professor Pickell | October 16th, 2023

Audience and Scope

This document is intended for the audience of computer science students who are looking to understand and create data mining algorithms that estimate housing costs using a database of house information. These algorithms can be very overwhelming when conceptualizing them because they are daunting projects. We can assume the computer science students have some understanding of the nuances of coding; however, they most likely will not have any understanding of machine learning and how to implement it. Attention will be given to best describing the process to allow the students to have a clear understanding of the topic.

Introduction

Machine Learning is a very useful technology that has thousands of applications that improve productivity exponentially. The first Machine learning programs were created in 1959 and weren't adopted into mainstream use until more recently in the 2020s. Machine learning in data mining involves algorithms that automatically learn patterns and relationships within large datasets without being explicitly programmed. These algorithms improve their performance over time by continuously refining their models based on new data. The goal is to extract meaningful insights, predict outcomes, or classify data based on these learned patterns.

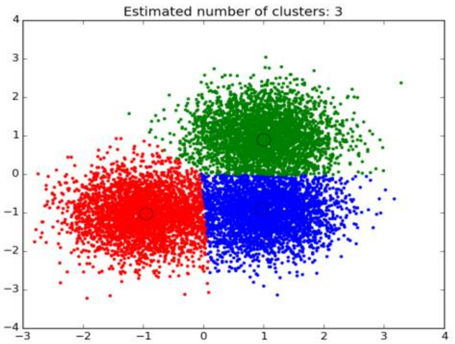
History of Data Mining

Data mining's processes began as early as computers became complex enough to hold datasets. In the 1960's, only a couple years after the creation of data mining, computer labs were full of human operators that were interacting with programs and every time the program would give an incorrect answer the operator would hit what they called a "goof" button. The Goof button is a button to cause the algorithm to re-evaluate incorrect decisions. The algorithm would then realize that the response would and the next time it would give a different answer until the operator approved of the response. After years of experiments and development the programs had become more and more accurate to the point where it could be relied on for correct math, coding answers.

Current State of Data Mining

Today's version of data mining is a lot simpler because almost everything has already been done in some form for reference. There are built-in functions and shortcuts in the different coding compilers to make the process of mining data way simpler. Coding compilers are a special program that translates a programming language's source code into machine code. Source code is using language that humans understand while machine code is a language that only computers understand. Programmers can take massive sets of data and write a couple of lines of code that interpret the data and perform different actions, such as clustering the data into separate groups as seen in the figure to the right.

Figure 1



How to Create the Algorithm

For this example, we will be using Python. Python is a programming compiler that has the most advanced machine-learning integration. The creation process can take anywhere from 3 to 20 hours depending on how in-depth your chosen data set is.

Identifying/Loading Dataset

A data set is a collection of related data that's intention is to be manipulated for learning purposes. Identifying the dataset can be one of the more difficult steps to this process. Once you have your topic you have to search the internet for a valid dataset to fit your needs. Not all datasets are created equal. During your search you want to identify a data set that is specific to your topic and has all the information you require within only one data set. The information should also be organized in a way that is compatible with your plan and from a reliable source.

When loading the data set into the compiler, depending on your computer brand you will have to use command line for Microsoft or terminal for Apple. Using your knowledge of the following software's you will have to navigate to where your CSV file is located on your computer.[3] A CSV file is a file type that is specific to holding datasets for coding purposes. Then using this directory, you will open your compiler by simply typing Python in the code line. After performing this action, the app version or browser version will automatically open up on your computer and you are able to begin coding using the data set you chose.

Data Processing

Data processing is a crucial step in any machine learning project. It involves cleaning, normalizing, and moving raw data into a format that can be used by special already designed machine learning algorithms. For housing cost estimation, this involves handling missing values, removing outliers, and converting categorical variables into numerical ones by using the "df.dropna()" and df["column"].unique() functions [5]. For our example of housing estimation, a variable like "neighborhood" will need to be changed into several columns showing whether the house is a neighborhood or not.

Feature Engineering

Feature engineering is all about getting the information from our dataset data. In the context of estimating housing costs, this involves creating new variables like "proximity to schools" or "average income of the neighborhood." These new features can provide additional insights that can be pivotal for the accuracy of the model. Like when a house's price might not just depend on its size but also on its location, age, and proximity to amenities.

Linear Regression Model

Linear regression is one of the most basic and widely used algorithms in machine learning. It tries to establish a linear relationship between the independent variables (features) and the dependent variable (house price). By fitting a linear equation to observed data, the algorithms can make predictions for new data. For housing costs, this involves predicting the price of a house based on features like its size, number of bedrooms, and location.

Random Forest Model

Random Forest is a more complex model that can capture non-linear relationships in the data. It works by making multiple decision trees during training and outputs the average prediction of the individual trees for regression problems. Decision trees are a type of machine learning used to categorize or make predictions based on how a previous set of questions were answered. In the context of housing costs, a Random Forest will consider a great number of decision paths based on the features of a house and use the combined knowledge of all trees to make a prediction. This causes an accurate model when there are complex interactions between features.

Conclusion

Estimating housing costs using data mining and machine learning is a constantly growing and evolving field. With the right dataset and by using powerful algorithms like Linear Regression and Random Forest, one can make accurate predictions that can be invaluable for new home buyers. As technology and data availability continue to grow, the accuracy and reliability of these estimates will only improve, offering more clarity and confidence to those trying to get into the real estate market.

References

[1] Zillow. "Zestimate: Home Values & Zestimate Accuracy | Zillow." Accessed on 20 October 2023. https://www.zillow.com/z/zestimate/

[2] Wikipedia. "Machine Learning." Last modified on 29 October 2023. https://en.wikipedia.org/wiki/Machine_learning

[3] NeuralNine. "House Price Prediction in Python - Full Machine Learning Project." Published on 25 November 2022. https://www.youtube.com/watch?v=Wqmtf9SA_kk

[4] Figure 1: Springboard. "Data Mining in Python: A Guide to Data Analysis and Data Mining." Published on 25 November 2022. https://www.springboard.com/blog/data-science/data-mining-python-tutorial/

[5] Lopamudra Nayak. "Dealing with Missing Data using Python." Published on 15 February 2022. https://medium.com/nerd-for-tech/dealing-with-missing-data-using-python-3fd785b77a05