Akinmerese, Oluwatobi, 2025, 13:6
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

International Journal of Science, Engineering and Technology
An Open Access Journal

House Price Prediction Model Using Data Mining

Akinmerese, Oluwatobi[1], Ifekandu, Chiamaka[2], Ezeoke, Evelyn[3], Osuji, Adaeze[4], Lawal, Esther[5]

[1,2,3,4]Department of Mathematical Sciences, Augustine University, Ilara-Epe, Lagos State, Nigeria.
[5]Department of Library and Information Science, Tai Solarin University of Education, Ogun State, Nigeria.

Abstract- Reliable methods for forecasting home prices are scarce in many areas, particularly in our nation. This research tackles this gap by applying machine learning techniques to anticipate house values based on key attributes. Using the Cross Industry Standard Process for Data Mining (CRISP-DM) framework as a reference, this research used the Linear Regression model. Data cleansing, feature selection and visualization were all part of the approach. It was found that accuracy increased when the dataset was transformed using logarithmic values and models were assessed using statistical techniques like p-values and the Bayesian Information Criterion (BIC). The results demonstrate that property prices may be accurately forecasted using a condensed dataset without sacrificing model performance.

Keywords: p-values, bayesian information criterion, CRISP-DM, data visualisation, linear regression, correlation, data cleaning and dataset.

I. INTRODUCTION

Finding hidden, legitimate, and maybe helpful trends in massive data sets is the goal of data mining. Finding unexpected or previously unidentified relationships in the data is the main goal of data mining. It is a multidisciplinary talent that makes use of database technologies, AI, statistics, and machine learning. Data mining yields insights that can be applied to scientific research, marketing, and fraud detection, among other areas. Data mining is also used as knowledge discovery, knowledge extraction, data/pattern analysis, information harvesting (Wikipedia, 2016).

The result of this prediction can be used by business firm and real estate agents to value homes in a particular area the tool will be able to give a predicted output benchmark price based on the characteristics of the home and by characteristics I mean things like how many rooms has or how much crime there is in an area plus a bunch of other factors the tool will be able to determine which factors are less important in determining the house price and what the premium is for living in a home where parents can send their kids to a good school in other words the valuation will be traceable

This will be able to bring together research on prediction markets to further their utilization by economic forecasters or real estate agents. That is why there is a need to predict the efficient house pricing for real estate customers with respect to their budgets and priorities. This research efficiently analyzing current house prices, thereby forecasting the future prices according to the user's requirements.

Statement Of The Problem

The general house characteristics are often listed separately from the asking price and overall description. Because these characteristics are separately listed in a specific way, that can be easily compared across the whole range of potential houses. Because every house also has its own unique characteristics, such as a particular view or type of sink, house sellers are able to provide a summary of all the important features of the house in description. Potential buyers can assess each of the real estate aspects listed, but because of the wide range of variables, it is practically impossible to provide an automatic comparison of all of them. This is also true in the other division house sellers have to make an estimation of the value based on its features in comparison to the current market prices of similar houses.

It is difficult to determine a suitable market price due to the variety of features.

Another major issues have to do with gathering data if there isn't sufficient data for our evaluation our prediction model may not be very precise enough

and since real world data is messy the three is going to be a lot of work in cleaning the data were we look out for missing data or incomplete data or errors and even bad formatting.

Aim And Objectives Of The Study

The aim of research work is to develop a computer application software that will be able to predict and estimate house prices as well as the contributing factors to the prices using supervised learning algorithms.

These are the specific objectives:

- To identify a suitable market price due to the variety of features.
- To develop a computer application program that will be able to mine knowledge from the dataset using supervised learning algorithms
- To evaluate the computer application program that will be able to mine knowledge from the dataset in comparison with other computer application programs using supervised learning algorithms

Methodology

This presents the procedures or methods employed in conducting this study and it includes: The Cross Industrial Standard for Data Mining (CRISP-DM). It is a problem-solving strategy which can adapt to the data mining process containing six stages. The stages are business understanding, data understanding, data preparation, modeling, evaluation and deployment. (Chapman,2000). Primarily, the data source was collected through oral interview from staff and agents in various estate firms. Data was analyzed and prepared, a suitable model selection was done, training and testing of the model was carried out and model evaluation was performed. The programming language that will be implemented in the course of this project research will by Python.

The tools used will be Numpy, Pandas, Scikit-learn etc. Python is relevant to this work because it is stable, flexible, simple and has the necessary tools needed

Significance Of The Study

This study will educate how the implementation of various algorithms can be used to visualize data and extract meaningful information. It will also educate on how the use of data mining techniques can be used in predicting different information based on the information given. This research will also serve as a resources base to other scholars and researchers interested in carrying out further research in this field subsequently the knowledge gained from this study will help understand roles of data scientist, data analysis, and does interested in machine learning, if applied will go to an extent to provide new explanation to the topic.

Scope And Limitation Of The Study

This study will cover the mode of operation of various models in data analysis and how it can be used to predict and valuate houses it will also include a graphical visualization of subsequent output.

The limitation focuses of the following constraints:

I. Financial constraints: the cost of sourcing for information and data that are involved in this work is high in the sense that we all know that information is money.

- Results from data mining may not be reliable if the data set is not diverse.
- Data mining needs large database which sometimes are difficult to manage.
- Overfitting: Due to small size training database, a model may not fit future states.
- Time

Definition Of Terms

- **Data:** Do you want to detect spam? Get samples of spam messages. Do you want to forecast stocks? Find the price history. Do you want to find out user preferences? When you Parse their activities on Facebook, the more diverse the data will be and the better the result. Large volumes of data in data mining is known as a dataset.
- **Features:** Also known as parameters or variables. Those could be car mileage, users gender, stock price, word frequently in the text. Stated differently, these are the elements that a machine should consider.

- **Algorithm**: an algorithm is a procedure or formula for solving a problem, based on conducting a sequence of specified actions.
- **Cost Function:** Often used by businesses to reduce costs and increase production efficiency, a cost function is a function of input prices and output quality whose values represent the cost of producing that output given those input prices.
- **Linear Regression:** Linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variable).
- **Correlation:** In other words, correlation is a measure of the degree of relationship between two variables. As a result, when one variable rises, the other rises, or when one variable falls, the other falls. The relationship between height and weight is an illustration of a positive correlation.
- **P-Value:** One type of inferential statistics is regression analysis. The P-value help determine whether the relationship that you observe in your sample also exist in larger population. Each independent variable's P-Value examines the null hypothesis, which states that there is no link between the two.

II. REVIEW OF RELATED LITERATURE

Introduction

In the 1960s, statisticians and economists used terms like data fishing or data dredging to refer to what they considered the bad practice of analyzing data without a priori hypothesis. In a 1983 paper that appeared in the Review of Economic Studies, economist Michael Lovell used the phrase "data mining" in a similarly negative manner. Lovell indicates that the practices "masquerades under a variety of aliases, ranging from "experimentation" (positive) to "fishing" or "snooping" (negative). Around 1990, the database community first used the term "data mining," which was typically associated with positive connotations. The term "database mining" was used briefly in the 1980s to promote a database mining workstation made by HNC, a San Diego-based firm, but it was later trademarked; researcher consequently turned to data mining.

Other terms used includes data archaeology, information harvesting, information discovery, knowledge extraction, etc Gregory Piatetsky-Shapiro coined the term "knowledge discovery in databases" for the first workshop on the same topic (KDD-1989) and this term became more popular in AI machine learning community. Nonetheless, the press and business groups began using the term "data mining" more frequently. These days, knowledge discovery and data mining are employed interchangeably. In the academic community, major forums for research started in 1995 when the first international conference on Data Mining and Knowledge Discovery (KDD-95) was started in Montreal under AAAI sponsorship.

Ramasamy uthurusamy and Usama Fayyad served as its co-chairs. A year later, in 1996, Usama Fayyad launched the journal by Kluwer called Data Mining and Knowledge Discovery as its founding editor-in-chief. Later he started the SIGKDD Explorations. The KDD international conference became the primary highest quality conference in data mining with an acceptance rate of research paper submissions below 18%. The journal titled Data Mining and Knowledge discovery is the primary research journal of the field. There is a huge amount of data available in the information industry. This data is of no use until it is converted into useful information. This massive amount of data must be analyzed in order to obtain relevant information.

Extraction of information is not the only process we need to perform. Other procedures including data cleansing, data integration, data transformation, pattern evaluation, and data display are also included in data mining. Once all these process are over we would be able to use this information in many applications such as Fraud Detection, Market Analysis, Production Control, Science Exploration, etc. Prediction and description are typically the two main objectives of data mining in practice. Prediction is the process of forecasting unknown or future values of other variables of interest using certain variables or fields in a data set. On the other hand,

description concentrates on identifying human-interpretable patterns that describe the data.

Consequently, data-mining efforts can be divided into two groups:

Using the available data set, descriptive data mining generates new, nontrivial information, or predictive data mining generates the model of the system described by the data set. On the predictive end of the spectrum, data mining aims to generate a model that can be used to carry out classification, prediction, estimation, and other related tasks. This model is written as executable code. On the other hand, descriptive, end of the spectrum, the goal is to gain an understanding of the analyzed system by uncovering patterns and relationship in large data set. The goals of prediction and description are achieved by using data-mining techniques:

Important concepts of house price prediction model

Supervised Learning

In a supervised learning setting, we use a labeled data that consists of features/variables and dependent variables (Y or response). This data is then fed to the learning algorithm that search for patterns, and a function that controls relationship between independent and dependent variables. The retrieved function may be then applied for the prediction of future observations.

Data Mining Prediction Model (Regression)

It is a data mining task of predicting the value of target (numerical variable) by building a model based on the one or more predictors the predictors can either be numerical or the categorical variables.

Linear Regression

Simple linear regression is a linear method used in statistics to model the connection between one or more explanatory variables and a scalar response (or dependent variables). For more than one explanatory variable, the process is called multiple linear regression.

Gradient Boosting

According to Ganjisaffar et al. (2011), gradient boosting can be applied to both classification and regression. Gradient boosting is a method for creating regression models that are made up of a group of regressors. It is an implementation of the concept for the data's regression predictor, after which the error residual is calculated. The main concepts entail we are making a set of predictions and finding the errors while reducing it.

Explanation of the existing system

The present system is not dunce proof and has certain drawbacks. Due of its manual nature, the current method has many potential drawbacks and weaknesses.

Some of them are: -

- **HUMAN RESOURCE:** the current system has too much manual work from filling a form to filling a document, delivering manifesto. This increases burden on workers but does not yield the result it should.
- **THORNY JOB:** Any changes to the current system would need more manual labor and be more prone to errors.
- **ERROR:** as the system is manages and maintained by the workers' error are some of the possibilities.

TABLES 2.5 REVIEW OF RELATED PROJECT WORK

Author & Year	Title of paper	Contribution	Problem observed	Gaps
Pagourtzi et al., (2003).	Journal of property investment and finance, 2003	The continuous expansion of data and record generation have become obvious as difficulties and challenges are more intricate with frequent	The problems observed in the estate firms include inadequate staff, untrained personnel, poor record keeping and time constraints.	(Pagourtzi et al., 2003).

		technological advancements		
Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani, (2017)	An introduction to statistical learning with R	Using fundamental statists and applying it with data mining techniques to produce accurate result	Network transforms data until they can be classified as a meaningful output	Complexity of processing requirements
Pradip Kumar Bala (2017)	Datamining for retail inventory management Xavier institute of management India.	Use of data mining techniques to discover customer performance and manage inventory accordingly	Gives the size of the coefficient of each independent variable which determines the direction or effect	Most models don't go well with certain situation hence you are required to test multiple models
Author K.B (2017)	Data mining for retail inventory management college of management India.	Using KDD methods such as clustering, k-means to uncover hidden trends common themes and patterns in big data	Use of data mining techniques to discover customer preferences and manage inventory accordingly.	Most existing methods are only capable of single item inventory.

III. RESEARCH METHODOLOGY

Introduction

House prediction is way of building up an assessment of significant worth for real property. Properties often require evaluation since each one is unique. The method for calculating a property's estimated monetary value is called house prediction. The value usually looked for is the property market price. It is a critical thinking process in which the impact of sociological, financial, physical and government forces is studied closely in relevancy to the property.

Over the past few years, a lot of analysts have highlighted the issue of improper or incorrect valuation and the uncertainty of the valuation. Examples of the known problems of estate firms include poor record keeping, insufficient staff, inexperienced agents, and time constraints the existing methods of valuation such as sales comparison approach, cost approach, income approach are still being challenged with; of improper or incorrect valuation and the uncertainty of the valuation procedure is the thing that every estate firm looks for this research work seeks to address the stated challenges posed by archaic methods of estate valuating and introducing new computerized methods of estate valuation having an efficient and reliable valuation prediction system is one of the great challenges faced by most of the estate companies, nevertheless, most of the estate firms admits having difficulties when it comes to estimating the worth of a property.

Regression is one of the machine learning algorithms that helps in prediction by learning through current existing data. With the definition, a property's value is determine based on parameters such as the land space, the geographical area, climatic conditions, economic and social factors, security and government. If we apply machine learning to these parameters, we would be able to determine values of property in a given geological zone (Chin et al., 2019).

ANALYSIS OF EXISTING SYSTEM

Some existing systems makes use of a linear approach to modelling the relationship between a scalar response (or dependent variable) and an explanatory variable (or independent variables). This case of one explanatory variables is called simple linear regression

Analysis Of Proposed System

This proposed system will make use of multiple features to determine the most accurate price for houses. This would make consumers understand every detail of how their money is being used because the model will use more attributes to determine the cost of a house. The proposed system for the project is by using JupyterNotebook, python, Bayesian information Criterion and scikitlearn to train the model using Linear Regression algorithm. JupyterNotebook has excellent built-in support for many data analysis and visualization routines, in particular, one of its most useful tools is that of effective exploratory data analysis, which is a natural fit in the context of data mining.

- The simplicity of the depiction makes the linear regression model appealing.
- Learning a linear regression model means estimating the value of the coefficients used in the representation with data that we have available.
- Multiple linear regression it's a new version of the linear regression which is considered to be more powerful which with the multiple variables or the multiple features it helps to predict the unknown value of the attribute from known value of the two or more attributes which will be also known as the predictors
- Schwarz's Bayesian information Criterion (BIC) is a model selection tool. The BIC score provides an estimate of the model's performance on a brand-new data set (testing set) after the model has been estimated on a specific data set (training set).

PROPOSED SYSTEM METHODOLOGY

CRISP-DM Methodology

The acronym CRISP-DM, which stands for Cross Industry Standard Process for Data Mining, is an open standard process model that outlines typical methods employed by data mining professionals for problem-solving techniques that are compatible with the six-phase DM process.

This process is based on six iterative main sectors which respectively are:

Business understanding
Data understanding
Data preparation
Modelling
Evaluation
Deployment

In other to develop the house price prediction model a series of task was taken which used the CRISP-DM

Methodology as follows: -

Knowledge of business: this model Although the project's title makes it clear, I came up with my question since, in the business knowledge phase, we usually want to know what our objective is for a given problem. For this one, it was to forecast home prices. Other examples would be forecasting the weather or a student's GPA. After understanding the business, we then collect pertinent data that will enable us to accomplish our objective.

Data understanding: after extracting data there are a number of things you should consider such as the type of data which normally comes in structural format, semi-structured, quasi- structured and unstructured practically examples would be (pdf, csv, xml, SQL-database etc.) after getting the data it would be important to evaluate the features of the database to know whether it conforms to our goal derived from the business understanding phrase.

Modelling: this step involves inputting the data into a specific algorithm in my case it'll be linear regression in other to uncover the relationships or patterns present in the dataset.

Evaluation: the final step of knowledge discovery from data is to verify that patterns produced by the data mining algorithms occur in the wider data set. Data mining methods do not always identify valid patterns.

It is common for data mining algorithms to find patterns in the training set which are not present in the general data set which leads to overfitting in the stage we'll split the data into training set and test set in other to measure the accuracy of the model built.

CHOICE OF TOOLS USED.

ENVIRONMENT

Anaconda: is a free and open-source distribution of the Python and R programming languages for scientific computing (data science, machine learning applications, large-scale data processing, predictive analytics, etc.) that aims to simplify package

management and deployment Jupyter Notebook: is an open-source web application that lets you create and distribute documents with narrative text, equations, live code, and visualizations. Data transformation and cleaning, machine learning, statistical modeling, numerical simulation, data visualization, and many other applications are among the many uses.

Python: it is an interpreted, object-oriented, high-level programming language with dynamic semantics. Python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python promotes code reuse and software modularity by supporting modules and packages.

PACKAGES AND MODULES

Sklearn

Numerous effective methods for statistical modeling and machine learning, such as dimensionality reduction, clustering, regression, and classification, are available in the sklearn library. This tool was used heavily in this project for running our regression analysis, splitting our data set into training and test data, loading dataset etc.

Pandas

Pandas is a software package designed for data analysis and manipulation in the Python programming language. In particular, it offers data structures and operation for manipulating numerical tables and time series. The three-clause BSD license governs the free software's release.

Matplotlib

This is a graphing library for Numpy, a numerical mathematics extension for the Python computer language. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter. This tool was used during the data visualization stage and helped in spotting outliers in the dataset and also used to show the regression line visually

Seaborn

Seaborn is a matplotlib-based Python data visualization package. It provides a high number points on the heatmap and clearly illustrated the correlation between out dependent variable and independent variable.

Numpy

A Python package called Numpy is used to work with arrays. It also has function for working in domain of linear algebra, Fourier transform, and matrices.

Statsmodels

It is a python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. This tool was used to find the variance inflation factor, Bayesian information Criterion and p-values

IV. IMPLEMENTATION AND RESULTS

Introduction

The simplicity of the depiction makes the linear regression model appealing. A particular collection of input values (x) is combined with a linear equation, the solution of which is the expected output for that set of input values (y). Consequently, the output value and the input values (x) are both numerical.

The linear equation assigns one scale factor to each input value or column, called a coefficient and represent by the capital Greek letter Beta (B). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimension plot) and is often called the intercept or the bias coefficient (Jason Brownlee, 2016).

Choice Of Tools

Anaconda: is free and open-source distribution of the Python and R programming language for scientific computing (data sciences, machine learning application, large-scale data processing, predictive analytics, etc.) that aims to simplify packages management and deployment

Jupyter Notebook: Using the open-source web program Jupyter Notebook, you can create and distribute documents with narrative prose, equations, live code, and visualizations. Data transformation and cleaning, machine learning, statistical modeling, numerical simulation, data

visualization, and many other applications are among the many uses.

Python: It is an interpreted, object-oriented, high-level programing language with dynamic semantics, python's simple, easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python promotes code reuse and software modularity by supporting modules and packages.

PACKAGES AND MODULES

- **Sklearn**

Classification, regression, clustering, and dimensionality reduction are just a few of the effective machine learning and statistical modeling algorithms available in the sklearn toolkit. This tool was used heavily in this paper for running our regression analysis, splitting our data set into training and test data, loading dataset etc.

- **Pandas**

Pandas is a data analysis and manipulation software library designed for the Python programming language. It provides procedures and data structures specifically for working with time series and numerical tables. The three-clause BSD license governs the free software's release. This tool was used to get the descriptive analysis of the dataset and to combine various dataset in other to make specific calculations.

- **Matplotlib**

This is a graphing library for NumPy, a numerical mathematics extension for the Python computer language. With the help of general-purpose GUI toolkits like Tkinter, it offers an object-oriented API for integrating plots into programs. This tool was used during the data visualization stage and helped in spotting outliers in the dataset and also used to show the regression line visually.

- **Seaborn**

Seaborn is a matplotlib-based Python data visualization package. It offers a sophisticated interface for creating visually appealing and educational statistical visuals. This tool was used to show the correlation metric in the dataset by plotting a huge number points on the heatmap and clearly illustrated the correlation between out dependent variable and independent variable.

- **Numpy**

A Python package called NumPy is used to work with arrays. Additionally, it provides functions for working with matrices, the Fourier transform, and linear algebra.

- **Statsmodels**

This Python module offers classes and functions for estimating a wide range of statistical models, performing statistical tests, and exploring statistical data. This tool was used to find the variance inflation factor, information criterion and p-values in other to determine.

System Requirements

System requirements are the required specifications a device must have in order to use certain systems effectively. These requirements include:

Hardware requirements

Software requirements

Hardware Requirements

Hardware requirements are the physical components the device must have in order for the system to work efficiently. The required hardware for the effective functioning of this system includes:

RAM: Minimum of 1GB

CPU: Pentium/133MHz processor or higher

Free hard disk space: 400MB

Sound card: Sound card that supports 16-bit recording

Microphone: Built-in microphone or an alternative microphone source (like a headset)

Internet connection: An internet connection is required

Software Requirements

Software requirements are the necessary software that has to be installed on the device before this system can work.

The only software requirement for this system is an operating system and the supported operating systems include: Microsoft Windows 7, 8.1 and 10 (32-bit and 64-bit).

System Implementation

Analysis of the Dataset

In this implementation we use a house pricing data based on Journal of environmental economics and management on hedonic house prices and the demand for clean air due to limited data and not enough features to make a proper study on data mining and under fitting issues faced while in my previous efforts the current data to carry out this study is of type bunch and it had various attributes chained to it which was further analyzed with python packages. The data tabulation offered information of the houses include: CRIME (per capita crime rate by town), PTRATIO (pupil-teacher ratio by town), LSTAT (percentage lower status of the population), RM (average number of rooms per dwelling), DIS (weighted distances to popular relaxation centres) e.t.c. total number of rows given was five hundred and six (506) and it has thirteen (13) columns which would be used to make the prediction based on after the transformation stage.

The steps in the analysis of the data set are as follows:

Extraction: Here, the dataset was extracted from a journal by Harrison, D. and Rubinfeld, D.L. Hedonic and it is maintained by Carnegie MELLON University. The data was a package inside the sklearn modules and it's of a type bunch data was broken down into arrays which was formatted to a data frame using pandas.

Data Exploration: During this stage, we get a general overview of what our data looks like. We thereafter observe all the dependent variables given the shape of the data, the shape of the data, the size of the data, we get a feel of the quantity of the data being we also take notes of the data points present in the dataset.

Data Cleaning: the data obtained from the repository is in the form bunch and had different objects chained to it one of which was the numpy.ndarray which was used to check the overall length of the attributes present in the dataset. The dependent and independent variables were extracted from the bunch file and inserted it into pandas dataframe which was used to carry further analysis using this package. Furthermore, the general idea of how the overall data looked like was gotten. It gives the ability to view the first five (5) rows and columns and the last five (5) with inbuilt functions.

Thus, data cleansing is an iterative procedure.

The next step was verifying the number of instances given in the data and checking for null values. The next step was detecting and correcting bad records before loading into the machine learning models. In other words, the data should be corrected in order to get the high accuracy

Data Visualization: Plotting a scatter of all the whole features gave an insight on the overall data so I could narrow down which features to pay more attention to Data visualization is the graphical representation of information and data. Using matplotlib in python I was able to visualize the given data using a bar chart to check for outliers and understand the distribution of the data set.

Detection Of Outliers: An outlier is an extremely high or extremely low-value value in the data it can be identified if whether the value is greater than interquartile Q3 + 1.5 or Q1 – 1.5 detecting the interquartile range is arrange the data in an order from the lower value to the higher value, now the mean is taken for the first set of values and the second set values now by subtracting both mean we can get the interquartile range the formula Q3 + (1.5)(quartile range) and for Q1-(1.5) quartile range and I have calculated using the python program. Checked for any inconsistent value that may have present if a label wasn't matched correctly by observing the various units given in the description of the data set.

Data Transformation: Applied a log transformation in other to fit the data into the linear regression model after applying this technique we got closer to the normal distribution which improved the general results of the model by minimizing our skew and getting a higher r-square for a normal distribution skew=0.

Training and Test Dataset: The dataset we made use of has been described in the beginning of this chapter taking the value of the independent variable which are the variables that help us make the prediction i.e. No of rooms, crime rate, accessibility to high ways e.t.c and splitting it into training set and also taking the same vale and splitting it into a test set they were dived into an 80/20 ratio which the parameter test_size=0.2 was used to get this split the method used here was built into sklearn module.

Implementation Process Algorithm

Linear regression can be straightforward simple linear approach for predicting a quantitative response Y on the basis of a single predictor variable X. it assumes that there is approximately a linear relationship between X and Y.

System Testing

After the implementation of the system has been done, the next thing to do is to test the system to ensure that it works as it is intended. The practice of testing an integrated system to make sure it satisfies requirements is known as system testing.

SYSTEM EVALUATION

- **Model Generation**

Performance metrics and Evaluation

The idea of a regression is to predict a real value which means number in regression model we can compute the several values the model most common terms are explained below

Coefficient of determination

The coefficient of determination R square summarize the explanatory power of the regression model and is computed from the sum of squares terms. The R square describes the proportion of variance of the dependent variable explained by the regression model and the equation is given below

V. SUMMARY, CONCLUSIONS AND RECOMMENDATION

Summary

In this research we were able to carry out house price prediction using Linear Regression we noticed that the algorithm can yield different result depending on the size of the dataset which was see when transforming the dataset into log prices and using various evaluation methods like p-values, VIF, BIC we were able to simplify our dataset without having to sacrifice the accuracy of the prediction model

Conclusion

This research will help in estimating the values of houses based on some explicit features that determine the value of the house

Recommendation

This research can be recommended to any real estate agent and anybody studying data mining or has interest in machine learning/ artificial intelligence the procedure carried out in this research is very detailed and easy for anyone to pickup

REFERENCES

1. Ethem Alpaydin (2020). Introduction to Machine Learning (Fourth ed.)
2. Friedman, Jerome H. (2016). "Data Mining and Statistics: What's the connection?". Computing Science and Statistics. 29 (1): 3-9.
3. Han, Kamber, Pei, Jaiwei, Micheline, Jian (2016). Data Mining: Concepts and Techniques (3rd ed.). Morgan Kaufmann
4. Michael I. Jordan (2014-09-10). "Statistics and machine learning". reddit. Retrieved 2014-10-01.
5. Resig, John. (16 March 2018.) A Framework for Mining Instant Messaging Services
6. Witten, lan H.; Frank, Eibe; Hall, Mark A. (2015). Data Mining: Practical Machine
7. Learning Tools and Techniques (3 ed.). Elsevier. Zimek, Arthur; Schubert, Erich (2017), Outlier Detection
8. Bergadano, 'F., Bertilone, R., Paolotti, D., & Ruffo, G. (2019). Learning Real Estate
9. Automated Valuation Models from Heterogeneous Data Sources. arXiv preprint arXiv:1909.00704.
10. Bitner, A., Król, K., Frosik, M., & Furczoń, M. (2020). Ecological Considerations in Real Estate Valuation. Journal of Ecological Engineering.
11. Bogin, A. N., & Shui, J. (2020). Appraisal Accuracy and Automated Valuation Models in Rural Areas. The Journal of Real Estate Finance and Economics, 60(1-2), 40-52.
12. Chin, W. M., Kit, N. L. W., & Fei, J. L. W. (2019, July). Valuation of Real Estate: A Multiple Regression Approach. In Proceedings of the 2019 2nd International Conference on Mathematics and Statistics (pp. 96-100).
13. Constantinescu, M. (2018). MACHINE-LEARNING REAL ESTATE VALUATION: NOT ONLY A DATA AFFAIR. Valuation Journal/Revista de Evaluare, 13(1).

14. Del Giudice, V., De Paola, P., & Cantisani, G. B. (2017). Valuation of real estate investments through Fuzzy Logic. Buildings, 7(1), 26.
15. Del Giudice, V., De Paola, P., Manganelli, B., & Forte, F. (2017). The monetary valuation of environmental externalities through the analysis of real estate prices. Sustainability, 9(2), 229.
16. Draper, D. W., & Findlay, M. C. (2014). Capital asset pricing and real estate valuation.
17. Real Estate Economics, 10(2), 152-183.
18. Gibson, Y. (2004) "Strategic property management" How can Local Authority Develop a Property Strategy"? Property Management, vol. 12 No. 3. pp. 9 - 14 MCB University Press, 0263-7472.
19. Masías, Víctor Hugo & Valle, Mauricio & Crespo, Fernando & Crespo, Ricardo & Vargas Schüler, Augusto & Laengle, Sigifredo. (2016). Property Valuation using Machine Learning Algorithms: A Study in a Metropolitan-Area of Chile.
20. Mooya, M. M. (2016). Real Estate Valuation Theory. Springer Books.
21. Nwuba, C.C. (2015) "Management of Property in a Depressed Economy", The Estate Surveyor and Valuer. Vol. 18, No. 2, Lagos
22. On, S. (2008). Mass appraisal of real property.
23. Pagourtzi, E., Assimakopoulos, V., Hatzichristos, T., & French, N. (2003). Real estate appraisal: a review of valuation methods. Journal of Property Investment & Finance.
24. Shiller, R. J., & Weiss, A. N. (2014). Evaluating real estate valuation systems. The Journal of Real Estate Finance and Economics, 18(2), 147-161.
25. Wang, K., & Wolverton, M. L. (Eds.). (2012). Real estate valuation theory (Vol. 8).
26. Springer Science & Business Media.
27. Wyman, D., Seldin, M., & Worzala, E. (2011). A new paradigm for real estate valuation.
28. Journal of Property Investment and Finance, 29(4-5), 341-358.
29. Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. Applied Soft Computing, 65, 260-271