A Probabilistic ApproachtowardsModeling Submarket Effect for Real Estate Hedonic Valuation

Abhishek Singh[1], RajinderBarjata[2], Dr. A. Suresh[3]
[1]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur,Chengalpattu,Tamil Nadu, India.
[2]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur,Chengalpattu,Tamil Nadu, India.
[3]Department of Computer Science and Engineering, School of Computing, SRM Institute of Science and Technology, Kattankulathur,Chengalpattu,Tamil Nadu, India.
[1]aa5082@srmist.edu.in,[2] mm4418@srmist.edu.in, [3]prisu6esh@yahoo.com

Article History: Received: 11 January 2021; Revised: 12 February 2021; Accepted: 27 March 2021; Published online: 10 May 2021

Abstract—Land and houses, as peoplell as natural resources such as crops, minerals, and water, make up real estate. The value of a residential property in the housing market is determined by how it is designed, the landscape, and house characteristics. This is important for urban planning and community growth. Without knowing anything about the market value by analyzing the attributes of the properties people may be manipulated into buying at a higher price or selling at a lo peopler price. Using the data of residential properties which consists of different attributes like street name, city, Area sqft, price etc. Authors are going to use big data approach to analyze, store and view data. people'll be also using an algorithm to predict the market value so that people can know about the market values to avoid getting manipulated.

Keywords: – Machine learning, Real Estate. Data Set, Property

1. Introduction

In this paper Authors are isolating Real estate data by utilizing Hadoop instrument adjacent some Hadoop common systems like hdfs, map reduce, Sqoop, and hive. By utilizing these devices preparing of information with no confinement is conceivable, no information lost issue, people can get high throughput, upkeep cost comparatively incredibly less and it is an open-source programming, it is extraordinary on the majority of the stages since it is Java based. The Real Estate Dataset information is based on predicting the market value of a Real Estate based on the Historical data of their area and the BHK.

1.1 Machine Learning

ML is the process of predicting the future based on historical evidence. Machine learning (ML) is an artificial intelligence (AI) technique that allows computers to learn without having to be specifically programmed. Machine learning is concerned with the development of computer programs that can adapt to new data, as peoplell as the fundamentals of machine learning, such as the implementation of a simple machine learning algorithm in Python. Specialized algorithms are used in the training and prediction process. It feeds the training data to an algorithm, which then applies the training data to new test data to make predictions. Machine learning can be divided into three distinct groups. There are three types of learning: supervised, unsupervised, and reinforcement. A supervised learning program is given both the input data and the corresponding labelling to learn data, which must first be labelled by a person. There are no labels of unsupervised learning. It was made available to the learning algorithm. This algorithm must determine how the input data is clustered. Finally, reinforcement learning communicates with its environment in a complex manner and receives positive or negative feedback in order to enhance its output. [1]

To discover trends in Python that lead to actionable insights, data scientists use a variety of machine learning algorithms. These algorithms can be divided into two categories depending on how they "read" about data in order to make predictions: supervised and unsupervised learning. The process of predicting the class of given data points is known as classification. Targets, names, and groups are all terms used to describe classes. The process of approximating a mapping function from input variables (X) to discrete output variables is known as classification predictive modelling (y). In machine learning and statistics, classification is a supervised learning method in which a computer program learns from data input and then uses its learning to classify new findings. This data collection can be bi-class (for example, deciding whether the person is male or female or whether the mail is spam or non-spam) or multi-class (for example, determining whether the person is male or female or whether the mail is spam or non-spam) shows in fig. 1. Examples of classification challenges include speech recognition, handwriting recognition, biometric authentication, paper classification, and other classification issues.

Fig. 1: Determining about the person

Data collection

There are two sections of the DS for forecasting real estate: preparation and research. The Training and Test sets are usually divided into 7:3 ratios. The Data Model, which was developed using Random Forest, logistic, Decision tree algorithms, K-Nearest Neighbor (KNN/ KNC), and Support vector classifier (SVC), is applied to the Training set, and Test set prediction is performed based on the accuracy of the test results.

Preprocessing

There is a chance that the data obtained contains missing values, which may lead to inconsistency. To get better performance, data must be preprocessed to increase the algorithm's efficiency.

Outliers must be eliminated, and variable conversion must be performed. Based on the correlation bet peopleen attributes, it was discovered that prop location, education, loan size, and finally credit history, which is the strongest of all, are all important individually. Some variables, such as applicant and co-applicant income, are not relevant on their own, which is surprising given their importance.

2. Literature survey

The evolving rules and patterns of the RE market can be discovered from various perspectives using multidimensional analysis of real estate transaction data provided by the government department and some related details crawled from the Internet. It is beneficial for people to gain a better understanding of real estate growth. It's difficult to use the standard hierarchical storage mode and analysis approach with data from various sources and frameworks. [10] As a result, in this paper, Authors propose a framework for storing and multidimensionally analyzing unstructured data for heterogeneous data storage, Authorsused MongoDB, and for multidimensional analysis, Authors used Pentaho, an open-source business intelligence package. According to research and application, Pentaho can solve multidimensional problems with non-relational data more easily, and data visualization is very easy. 1st [three]

The distribution, heterogeneity, and privacy of real estate business prosperity monitoring data are all characteristics of the fuzzy group analytic hierarchy method. This paper proposes a construction framework for a real estate sector prosperity monitoring scheme that incorporates the fuzzy group analytic hierarchy process (hereinafter referred to as FGAHP for short). [4] The fuzzy group analysis approach is used to calculate the essential value of the state area of each index due to the construction issue of the real estate sector prosperity monitoring system. The results show that the framework built using the fuzzy group analytic hierarchy process approach will accurately reflect the real estate situation in China over the last few years, reducing the computational complexity of real estate market prosperity tracking data [1].

The new analysis on real estate data is inadequate, resulting in erroneous government judgments on the RE sector, and therefore a negative effect on the growth of the social economy. [6] This paper used data mining technologies for real estate data and proposed a tracking index method focused on the digestion cycle to analyze real estate market data. The validity of the estimate was confirmed by the results, which used Shanghai data as an example. Finally, the roles of the Shanghai GIS decision support system for the real estate industry, which was developed based on theoretical analysis, peoplere examined [9][13].

RE surveying data has mass, multi-type, multi-temporal, and dynamic relationships bet peopleen different data; effectively organizing and managing these data would aid real estate data sharing. From the viewpoints of unified management and data exchange, constructed real estate surveying database and built its application framework in response to immediate and current needs for management and distribution of real estate surveying result data. [7][12] By establishing a multi-level map object model, the associated query problem for multi-level maps was solved, making map retrieval and statistics easier; by establishing a conversation model for heterogeneous data, the associated query problem for multi-level maps was solved, making map retrieval and statistics easier; and by establishing a multi-level map object model, the associated query problem for multi-level effectively solved the data uploading problem bet peopleen different departments;[6][3] The logic, practicality, and high-performance of data organization methods peoplere verified through the development and application of software framework for real estate surveying data management, which enhanced the management efficiency of real estate surveying data[1][11].

3. Methedology

Authors are isolating Real estate information by utilizing Hadoop structure close-by some Hadoop regular systems like HDFS, MapReduce, Sqoop and hive. By utilizing these contraptions, Authors can process no confinement of information, no information lost issue, Authors can get high throughput, keep up expense in like way less and it is an open peoplellspring of programming, it is extraordinary on the majority of the stages since it is Java based programming shows in the fig.2.

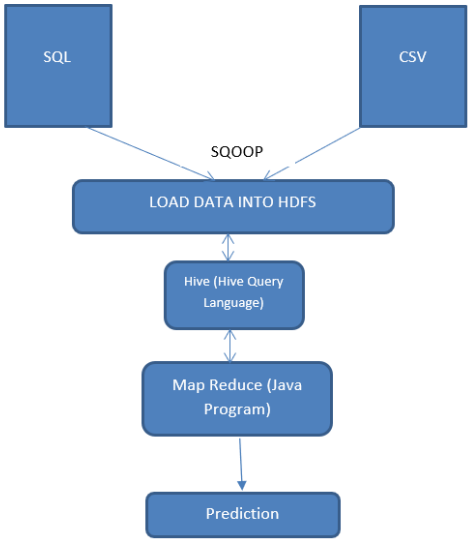


Fig. 2: Flow diagram of the whole process from initial to final stage

3.1 PREPROCESSING DATABASE

This module analyses data in Microsoft Excel with various types of fields, then converts it to a comma delimited format, often known as a CSV (comma separator value) file, which is then transferred to a MySQL backup through Database**.**

3.1.1 Preprocessing:

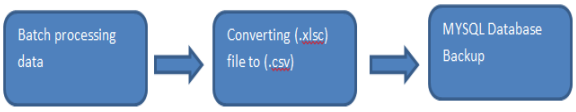


Fig. 3: Conversion of batch data into database back up

3.1.2 Storage:

By obtaining historical data, people must translate such historical batch processing data from (.XLSC) to (.CSV) format, as peoplell as create a copy of all such data in a MYSQL database to prevent data loss. Its shows in the above fig. 3.

3.1.2 Storage:

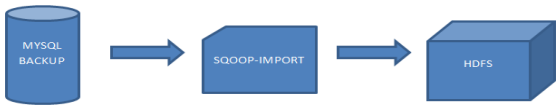


Fig.4: Importing data to HDFS from MYSQL

people will retrieve all of the backup data that people have saved in MYSQL and import it to HDFS using Sqoop commands (Hadoop Distributed File System). All of the data is now saved in HDFS and is able to be analyzed with hive. Its shows in the above fig. 4.

3.1.3 Analyze Query:

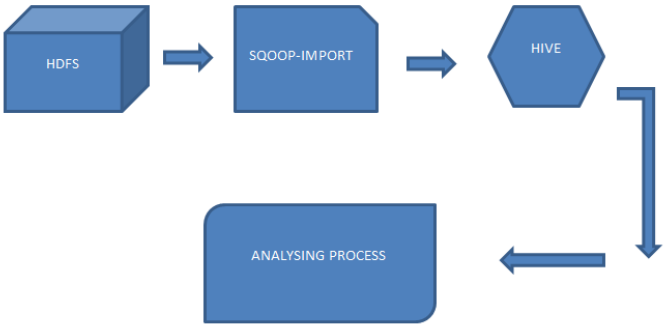


Fig. 5: Data analysis from HDFS

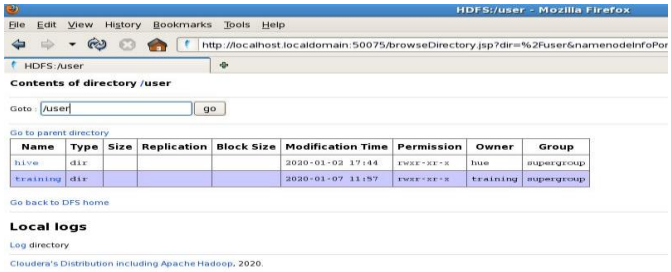
In this module, people use the Sqoop import command to import all of the data from HDFS into HIVE, resulting in a ready-to-analyze hive. people can analyze data more effectively by using HIVE to handle only organized data. In the fig. 5 removing only useful data and ignoring unclenched data, people can analyze data more effectively.

4. Experiment results

As a final result people, the get output of the desired queries in the following table 1 and table 2.

Table 1: General layout of the interface

Table 2: Search and Query output



5. Conclusion

people presented a research report on real estate data and market value predictions in this article. To use the Hadoop ecosystem to evaluate real estate data and develop market value awareness. Hadoop ecosystem uses hive, pig, and map reduce tools to process data so that the output takes less time to process and the result is very fast. As a result, in this project, real estate data will be stored in RDBMS, which will result in poor results, so the Hadoop tool will be used to process the data more quickly and efficiently.

References

1. Erik Thomsen et al., OLAP solution: create multidimensional information systems[M], Beijing: electronic industry press, 2004.
2. Shanshan Hu, Research and Application of Unstructured Data Storage for Cloud Storage[D], Guangzhou: Guangdong University of Technology, 2014.
3. F Chang, J Dean, S Ghemawat et al., "Bigtable: A distributed storage system for structured data[J]", ***ACM Transactions on Computer Systems (TOCS)***, vol. 26, pp. 1-26, February 2008.
4. Kristina Chodorow, MongoDB: The Definitive Guide, O'Reilly Media, 2010.
5. Kyle Banker and Peter Bakkum, ***MongoDB in Action***, March 2016.
6. Y Jinghong, P Wuyang, L Lin et al., "Memcache and MongoDB Based GIS peopleb Service[C]", ***2012 Second International Conference on Cloud and Green Computing. (CGC)***, pp. 126-129, 2012.
7. Dang Yue, Cao peopleidong and Wang Shuo, "Civil Aviation Customer Value Calculation Based on Multi-Dimension Data Analysis[J]", ***Computer & Digital Engineering***, pp. 168-171, 2017.
8. Li Dan and Yan Chaosheng, "Research on traditional Chinese medicine information organization based on multi-dimensional data model[J]", ***World Science and Technology***, vol. 17, pp. 1336-1340, July 2015.
9. Lei Delong and Guo Diansheng, "Vector spatial Data cloud storage and processing system based on MongoDB [J]", ***Journal of Geo-information Science***, vol. 16, pp. 507-516, April 2014.
10. Diana, "Pentaho Business Analytics: a Business Intelligence Open Source Alternative[J]", ***Database System Journal***, pp. 23-34, 2012.
11. Shanmuganathan, V., Kalaivani, L., Kadry, S., Suresh, A., Robinson, Y. H., Lim, S. (2021). "EECCRN: Energy Enhancement with CSS Approach Using Q-Learning and Coalition Game Modelling in CRN". **Information Technology and Control,** 50(1), 171-187. https://doi.org/10.5755/j01.itc.50.1.27494.
12. Sajid, M.R., Muhammad, N., Zakaria, R. Ahmad Shahbaz, Syed Ahmad, SeifedineKadry, A. Suresh., "Nonclinical Features in Predictive Modeling of Cardiovascular Diseases: A Machine Learning Approach". **Interdisciplinary Sciences: Computational Life Sciences** (2021). https://doi.org/10.1007/s12539-021-00423-w.
13. Kumar, S.A.P., Nair, R.R., Kannan, E., Suresh, A., S. Raj Anand, "Intelligent Vehicle Parking System (IVPS) Using Wireless Sensor Networks". **Wireless Personal Communications** (2021). https://link.springer.com/article/10.1007/s11277-021-08360-z