House Prices

Anusha[1], Vandana Kolli[2]

Abstract
In this paper, we analyzed the house price based on multiple features using data mining algorithms. In order to select a prediction method different regression methods were explored and compared. The dataset we used describes the sale of individual residential property in Ames, Iowa from 2006 to 2010. We built models using linear regression, decision trees and random forests and compared their performances. The problem is how to reduce multiple features to a minimal number(feature extraction) that can completely predict the target price variable. The results show that data mining algorithms produces high prediction accuracy in house price analysis and prediction.

Keywords
house — price — prediction

[1] *Computer Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
[2] *Computer Science, School of Informatics and Computing, Indiana University, Bloomington, IN, USA*
***Corresponding author**: todo

Contents

1. Introduction

1.1 Problems in House Price Prediction

Real estate properties will always be a safe option of investment for a common person. It is an investment which does not decline in value rapidly. The rapidly growing real estate industry has become an important part of national economy. Problems such as rapid raise of house price, high vacancy ratio require the Government to institute industrial policies to help industry development. With the increase on the data, the analysts need to apply good tools and predict house prices to help consumers. There is a lot of house data available with different features like Lot Area, Street, neighborhood etc. The important thing is to find a efficient data analysis tool to transform the data into knowledge and information[1].
The data comprises of the multiple features of the house which can be considered by potential buyers , but it is impossible to provide an automated comparison on all the features as they are diverse. The diversity of features makes it challenging to predict correct market price. Feature Extraction is one of the biggest challenges faced with data having large feature set. The house predict

1.2 Data Minning on House Price prediction

Data mining is the process of understanding useful and structural patterns in data, thus referring to the overall procedure of information discovery from data. These data specific discoveries became popular in the domain of machine learning and artificial intelligence. Some of the most popular machine learning algorithm include - decision trees, linear regression, logistic regression, neural networks, random forests and so on. Each algorithm suits some problems better than others. Data mining techniques are used to extract knowledge from data using a different of methods that are divided in two groups: classification and regression. Classification learns to map, or classifies, an item in the data into one of several predefined classes while regressions maps the item to a continuous numerical prediction.

Data mining draws inferences from data to understand the patterns of correlation among data values and to predict the future data values. By calculating the differences between the actual and predicted price as an error, we compare the results and accuracy of the predicated prices by the Decision Trees and Neural Networks algorithms.

1.3 Interests and Motivation

TODO The consumers of the house price prediction can be - homeowners, investors, tax assessment department, insurers etc. These consumers should get an accurate prediction of the final price of house. It is interesting to learn based diverse features on the existing consumer data and predict accurate house prices for any new data.

[1].

2. Background

In this paper we used machine learning algorithms like linear regression, decision trees and neural networks to predict the house prices. Machine learning techniques can be divided into supervised and unsupervised. Our analysis takes into consideration the provided target variables in making predictions, thus it is supervised. In this section we will discuss about these techniques in detail.

2.1 Linear Regression

A Regression Model defines three types of regression models: linear, polynomial, and logistic regression. Linear regression, the most widely used of all statistical techniques, allows to summarize and study relations between two continuous variables[2]. It attempts to model the relationship between these two variables by fitting a linear equation to observed data.
Let Y be the target value which we need to predict.
Let X be the independent variables from which target is predicted, then
The linear regression line has an equation of the form:

$$\text{Y = a + bX}$$

Mos common method to fit a regression line is the method of least squares which calculates the best fitting line for observed data by minimizing the sum of squares of the vertical deviations from each data point to the line.

2.2 Decision Trees

A decision tree builds regression or classification models in the form of a tree structure. The tree includes root node, branches and leaf nodes. Each internal node denotes a test on an attribute, each branch denotes the outcome of a test and each leaf node holds a class label. Decision trees can handle both categorical and numerical data.
Different algorithms and impurity measures are used for building a Decision Tree. In our analysis we have used CART(Classifica and Regression Tree) algorithm. In our paper since we are dealing with prediction of a continuous variable(price), the decision tree is called regression tree.

Gini Index

It measures the inequality among values of a frequency distribution. Gini index of 0 means perfect equality(where all values are same) and Gini index of 1 means maximum inequality among values. The impurity(or purity) measure used in building decision tree in CART is called GINI index.

$$GiniIndex = 1 - \sum_j p_j^2$$

2.3 Neural Networks

Artificial Neural network is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.

3. Data Analysis

TODO

3.1 Subsection

TODO

Table 1. Table of Grades

Name		
First name	Last Name	Grade
John	Doe	7.5
Richard	Miles	2

4. Problem

TODO

5. Experiments Data

TODO The model proposed in this paper was trained on a very large and diverse dataset. In this section we describe the details of the dataset. In addition, we also discuss the details of the various standard techniques that have been used for the problem, with which the performance of the model was compared. The results of the various techniques are given in the next section.

6. Results

7. Summary conclusion

8. Future Work

In future we are planning to work with random forests. In a random forest, each node is split using the best among a subset of predictors randomly chosen at that node. This somewhat counter intuitive strategy turns out to perform very well compared to many other classifiers.

[1] And some mathematics $\cos\pi = -1$ and α in the text.

Figure 1. Wide Picture

9. Appendix

Acknowledgments

So long and thanks for all the fish [**?**].

[1] https://www.irbnet.de/daten/iconda/CIB5807.pdf [2] https://onlinecourses.science.psu.edu/stat501/node/251