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

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.

[1].

2. Background

A database is software application that allows a large collection of data. It allows fast and optimal retrieval of data. In Bayesian terms, Probability signifies the belief that one has on the outcome of an event.

Given a sample space Ω, Probability of an advertisement $P(ad_c licked)$ being clicked, is number of times the advertisement is clicked $N(ad_c licked = 1)$ by the total number of times the advertisement appears, i.e. $N(ad_c licked = 0) N(ad_c licked = 1)$.

$$P(ad_c licked) = \frac{N(ad_c licked=1)}{N(ad_c licked=0)+N(ad_c licked=1)}$$

A clustering algorithm is used to identify and correlate similar data points. The K-Nearest Neighbor algorithm, identifies the k nearest neighbors for a given data point(a single row in a dataset). The *KNN* algorithm identifies similar data points using a distance metric. TODO - A distance metric is fill in from assignment 1

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

5. Experiments Data

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. The dataset provided is a sample of user's page views and clicks. The data is from multiple publisher sites from United States between 14-June-2016 and 28-June-2016

6. Results

7. Summary conclusion

8. Future Work

9. Appendix

Acknowledgments

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

[1] https://www.irbnet.de/daten/iconda/CIB5807.pdf

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

Figure 1. Wide Picture