Analysis and Prediction of House Prices for the City of Ames

Anusha Ramamurthy[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] *Data 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 for any new data.

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:

[1] $$Y = a + bX$$

Most 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 (Classification 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

3.1 Data Description

The dataset consisted of information about 2919 houses in the City of Ames, Iowa, United States of America. These houses were split into two sets, one with information about 1460 houses, each has 81 attributes, 79 of them describe various aspects and quality of the house, two others being the ID, identification number for each house and the other is an attribute we wish to predict, the Sale Price of the house. Among the 79, we found 23 attributes to be nominal, 23 to be ordinal, 14 were discrete and 19 continuous set of values [3].

Table 1. Table of Attribute Types

Attribute Type	Count
Nominal	23
Ordinal	23
Discrete	14
Continous	19

All the continuous variables are the dimensions measured in square feet, for example the size of the lot, the basement area, the size of the living area and so on. The 14 discrete variables are the count of the various commodities present in the house, for example the count of the bedrooms, bathrooms and so on. The 46 categorical variables describe aspects like the quality of the house, fireplace pool and so on, as well as information about the neighborhood the house belongs to. Hence we couldn't eliminate missing values as there would

[1] where a and b are constants.

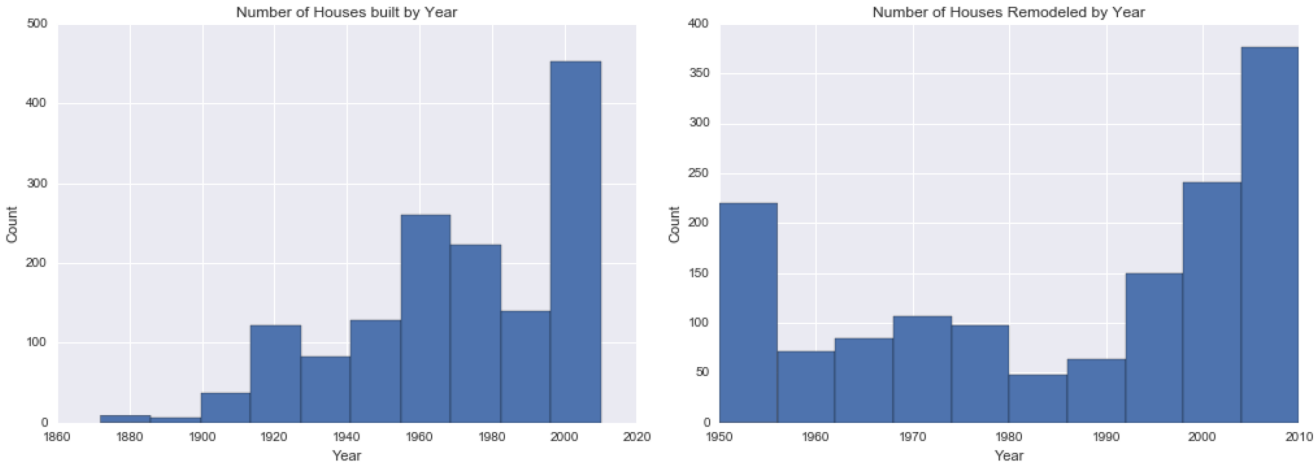

Figure 1. Count of Houses Built and Remodeled

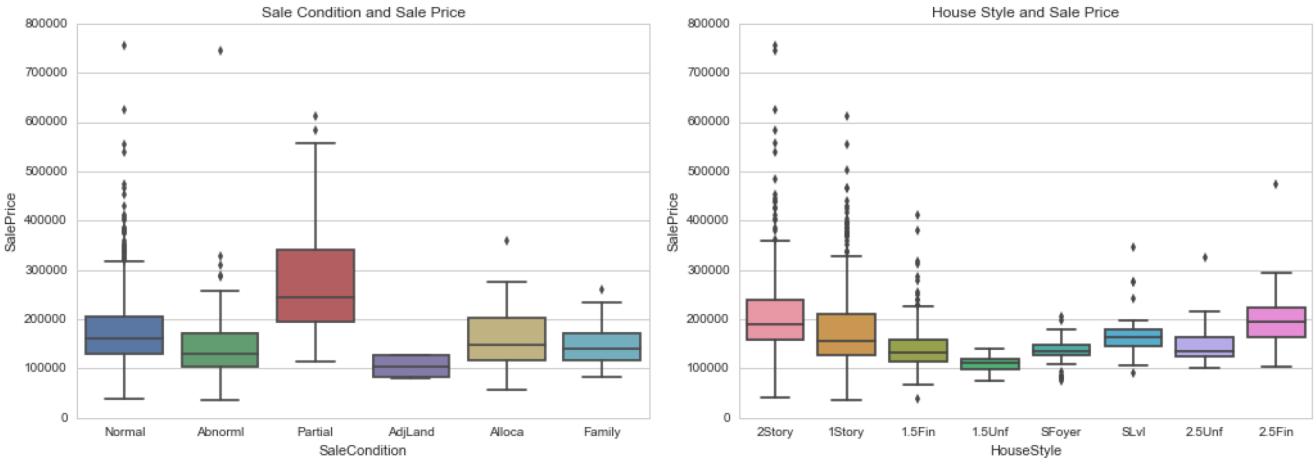

Figure 2. Distribution of Sale Price

be no row available.

3.2 Exploratory Analysis

We begin our analysis by checking for duplicates, a check of Id values showed 1460 unique observations. We found that among the 81 attributes, 19 of them had missing values. An initial exploratory analysis of the 62 attributes that had no missing values revealed some interesting insights.

A thousand four hundred and sixty houses were sold in just four years. The year 2009 appears to have seen the highest number of sales with 338 houses sold. Around 551 of the house's appear to have been built between 1999 and 2010, while 909 of the houses that were sold between 2006 and 2010, were built between 1870 and 1989. Around 178 of these houses were remodeled in the year 1950, 20 of these were built in the year 1920. Clearly the year a house was built seems to be a factor in the sale of a house, but only further investigations can say if it was an advantage or not. Does the age of house make it more attractive to a buyer? It appears hard to say. From the graphs, most of the houses appear to have been remodeled during 1950-1955 and 2000-2010.

It also appears that amongst most forms of Sale, Normal conditions of the sale appear to have fetched the highest price. But, houses that were not complete when last assessed also seem to fetch a higher price(Partial Condition). In fact looking at the distribution it appears that these fetched a broader range of prices than most other form of Sale Conditions. It also appears that Two story houses fetched the highest price as well as having Two and one-half story and One story houses fetched the next best prices. In both graphs we see houses that fetched a very high price, these may be possible outliers.

3.3 Missing Values

A large part of analysis depends on the amount of data we have. A large volume of data allows our data mining algorithms to train better and eventually reduce the margin of error. With the current data set we found 19 attributes containing missing values. Every row of our observations had at least one attribute with a missing value.

Table 2. Table of Missing Values

Attribute Name	Type	Missing
Electrical	Nominal	1
MasVnrType	Nominal	8
MasVnrArea	Continous	8
BsmtQual	Nominal	37
BsmtCond	Nominal	37
BsmtExposure	Nominal	38
BsmtFinType1	Nominal	37
BsmtFinType2	Nominal	38
GarageType	Nominal	81
GarageYrBlt	Continuous	81
GarageFinish	Nominal	81
GarageQual	Nominal	81
GarageCond	Nominal	81
LotFrontage	Continuous	259
FireplaceQu	Nominal	690
Fence	Nominal	1179
Alley	Nominal	1369
MiscFeature	Nominal	1406
PoolQC	Nominal	1453

3.4 Replacing Missing Values

Looking at each of the missing attributes we had to find a way to fill them without loss of existing relationships and natural distribution of the data. Some of the existing methods to replace missing values is by filling them with the mean , median or the mode. Since 16 of our missing values are Nominal data types, mean and median are meaningless. If we choose to replace them by the mode blindly, we might make our data skewed. Hence we decided to look at each attribute of the house, before we fill the missing values.

For the Electrical Attribute, it could take the four possible values. We found that 'FuseA' appeared to be the most common type of electrical type across homes of in the 'Timber' Neighborhood and replaced our missing value with the same. Attributes Masonry veneer type and Masonry veneer area appeared to be linked. For values of Masonry veneer type as 'None', the Masonry veneer area always had 0.0 square feet as the value. For those observations where Masonry veneer area alone was missing, the Masonry veneer type was also missing. Hence we filled Masonry veneer type as 'None' and Masonry veneer area with values '0.0'.

Attributes 'BsmtQual', 'BsmtCond', 'BsmtExposure', 'BsmtFinType1', 'BsmtFinType2' also appear to be related as they describe properties of the basement area of the house. We found that these attributes that were missing for the observations appeared to be houses with no basements. On analysis we found that the BsmtFinSF1, BsmtUnfSF, TotalBsmtSF which were values about the dimensions of the basement were all zero. Hence we filled each of the missing values with 'NA', which conveys the information that no basement is present.

Attributes 'GarageType','GarageFinish, 'GarageQual', 'GarageCond', 'GarageYrBlt' give description about the garage of the house. These values appear to be missing for houses that had no garage. Hence we filled each of these values with NA, except the attribute 'GarageYrBlt' which we felt was not as informative if there is no garage in the first place. For most houses the 'GarageYrBlt' appears to be the year the house was built. Hence we choose to remove this attribute all together. For the attribute 'FireplaceQu' we have no other correlated attribute that could give us an intuition on how to fill its missing values. We decided to assume that the missing values signify a house without a fire place, as most houses in Ames appear to have Central heating and filled it with 'NA'. We did the same for attributes 'PoolQC' and 'MiscFeature' as they appear to be for housings without a pool or any other miscellaneous feature like elevators. For the attribute 'Fence', we assumed that most houses would have the bare minimum fencing and replaced it with 'MnWw' which stands for minimum wood or wire fencing. For the attribute 'Alley' we choose to fill this attribute with 'NA' after a study of the document on Housing Information, on the City of Ames website which stated that attribute 'LotFrontage' and 'Alley' need not have a minimum value[4].

4. Problem

We now have a complete dataset, with 78 attributes that could help us in determining the price of the house. We choose to first build a linear model to predict the Sale Price. This was motivated by the paper on this dataset described in [3]. This problem was and is assumed to be a best solved by regression models. It is intuitive to think that each extra bedroom would cost x amount of cash, so we assume that the independent components of the house linearly influence the price of the house. For example

$$Cost_{(house_{(i)})} = (Number_{(bedrooms)} * Area_{(bedroom)} + Number_{(bathrooms)} Area_{(bathroom)} + Number_{(kitchens)} * Area_{(kitchen)}) * Cost_{(persqfeet)}$$

However, most of us while looking for a house, either like one aspect of the house more than anything else. So for some person X, the best think is a porch, and the price X would pay for it is not a tangible value. We have no way of quantifying a likable attribute. Hence in our opinion, a linear equation for such a problem has its own caveats. What we needed was an algorithm that will look at all the existing training samples and learn the distribution without making prior opinions about the parametric nature of the problem.

Hence we choose to next apply a non parametric algorithm. Decision trees are popular algorithms that create branching by asking yes or no questions. CART or Classification and Regression Trees are extensively used in Predictive Modeling. A regression tree partitions each observation such that those with same values are grouped together. Hence it maintains the natural grouping of the data. A tree produced by the algorithm

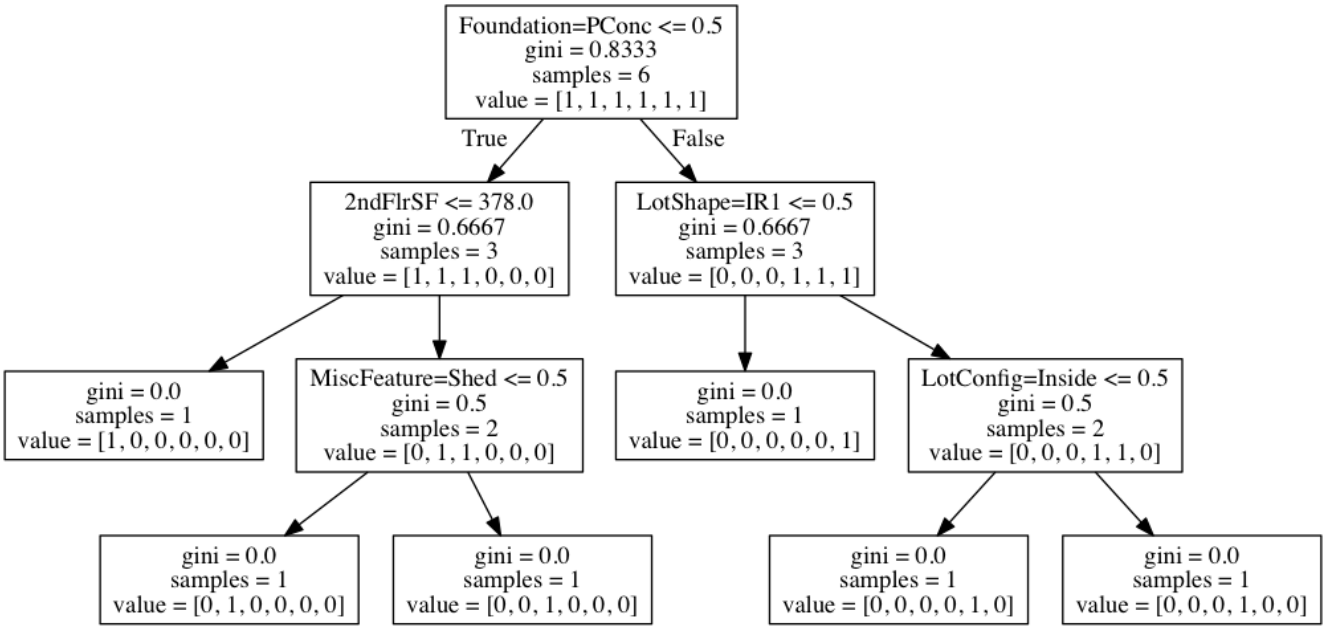

Figure 3. Decision Tree for 10 observations

for small set of 10 observations is included.

A Decision Tree can over fit and not generalize well. Hence it fails to perform well on new observations. One method to over come the bias and variation is by using Boosting [5]. A set of weak decision trees are combined with adjusted weights[6]. Adaptive boosting or Adaboost implementation was hence used to generate better results.

5. Experiments

The data was present as a competition in the Kaggle website[7]. A detailed description of the attributes is also present. The data set is split into two files, one with the Sale Price given, the other without. We used the data file with the Sale Price present to implement and test our hypothesis. We began by doing some initials checks on the data set, like the size of the data set, the attributes. We then moved to the cleaning and preprocessing of the dataset. This involved the process of finding and replacing the missing values.

All of our implementation was using Python programming. For most regression algorithms, the data has to be of Real values, namely continuous data types. Hence we had to use a method of changing our nominal data types to real valued numbers[8].

We split our Data set into 70-30 ratio. 70% of our data is split to train the algorithm and 30% is used in testing. We applied Linear Regression, Decision Tree and AdaBoost for Decision Tree Regressors We also applied K-fold Cross Validation methods for each algorithm[9].

Table 3. Table of Implementation Specifications

Name	Type
Operating System	MAC Sierra
RAM Size	16 GB
Programming Language	Python
Version	2.7
Editor	Ipython Notebook
Libraries	pandas, scikit-learn, NumPy matplotlib,pydotplus,graphviz

6. Results

We trained our algorithms with 78 attributes, to predict the House Prices. Linear regression yielded extremely poor results. We then applied Decision Tree Regressor For the 70 training,30 test split it yielded a score 61%. On K-fold cross validation with $k = 10$ Decision trees gave better score of 76%.

We then applied AdaBoost for Decision Tree Regressor For the 70 training,30 test split it yielded a score 85%. On K-fold cross validation with $k = 10$ AdaBoost gave an average score of 85%.

From the graphs, we do se several outliers that appear to be never accurately predicted by either of the algorithms. It does appear that Decision Tree comes very close to predicted the Sale Prices, but never quite gets there. The predicted and actual prices appear to be more closer in the case of Adaboost. The Pearson Coefficient Ratio for Decision tree for Actual and Predicted values was around 0.86, while for Adaboost it was 0.93.

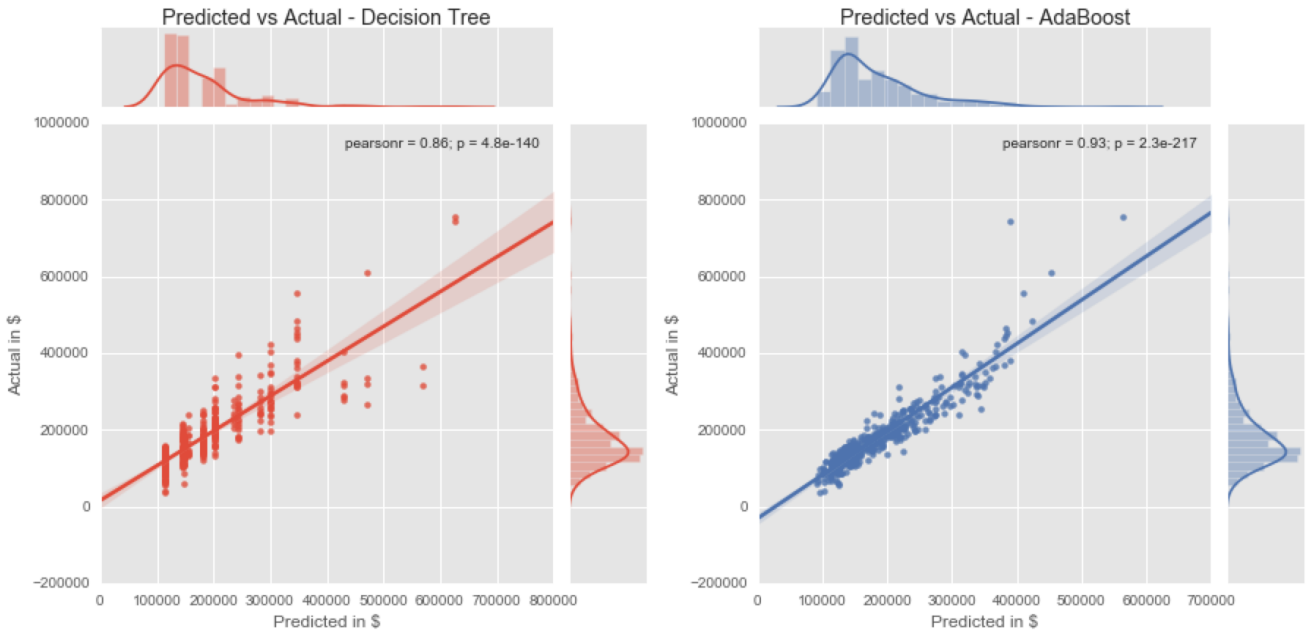

Figure 4. Results

7. Summary conclusion

The results from Linear regression were disappointing, but in a way it makes sense. There are a number of features that contribute towards making the decision to buy a home and they may not all be linearly correlated. While decision trees were able to learn better, we would like to explore other algorithms and approaches to get a much better accuracy of Sale prices.

8. Future Work

In future we are planning to work with Random forests and Lasso Regression. We would also like to explore PCA and other dimensionality reduction techniques. An idea we would like to explore is combining this dataset with the Boston Housing Prices Dataset to analyze if Housing Sales follow similar or different patterns across cities.

9. Appendix

Acknowledgments

We would like to thank Prof. Mehmet Dalkilic for giving us the opportunity and resources to work on this project. We would also like to thank Hasan Kurban and Kurt Zimmer for all their help.

References

[1] Qi Ming LI Xian Guang LI. The application of data mining technology in real estate market prediction.

[2] Stat 501 online course.

[3] Dean De Cock. Ames, iowa: Alternative to the boston housing data as an end of semester regression project. *Journal of Statistics Education*, 19(3), 2011.

[4] Iowa City Council of City of Ames. Ordinance 4252.

[5] Robert E. Schapire. The strength of weak learnability. *Machine Learning*, 197-22(5), 1990.

[6] Robert E Schapire Yoav Freund. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55:119–139, Aug 1997.

[7] Kaggle. House prices: Advanced regression techniques.

[8] John Langford Alex Smola Josh Attenberg Kilian Weinberger, Anirban Dasgupta. Feature hashing for large scale multitask learning. *Yahoo! Research*.

[9] PhilipSpector LeoBreiman. Submodel selection and evaluation in regression- the x-random case. *UniversityofCalifornia , Technical ReportNo. 197*, March 1989.