Land Price Prediction using Machine Learning Algorithm

Shubham Singh[1], Monika Nag K J[2]

[1]Dept. of Information Science and Engineering, NIE, Mysore
[2]Dept. of Information Science and Engineering, NIE, Mysore

---***---

Abstract -*The development of a model which can predict house prices can assist a house seller, buyer or real estate agent to make better, informed decisions based on current price valuation. Housing prices are increasing rapidly, yet the numerous websites online where houses are sold or rented are less likely to be updated on a regular basis. In various cases, individuals interested in selling a house or apartment might include it in some online listing, and forget about updating the price. In this paper, we aim at developing a machine learning application that takes into consideration the various factors in the real estate market in real time, i.e. houses that are listed with a price substantially below the market price, and factors such as demography etc. To predict the property prices, we ensemble two different ML architectures, based on Random Forest(RF) and Linear Regression(LR).*

Key Words: Price, Machine Learning, Linear Regression, Random Forest, Real Estate, Data etc.

1. INTRODUCTION

A housing market is any market for properties which are negotiated either directly from their owners to buyers, or through the services of real state brokers. People and companies are drawn to this market, which presents many profit opportunities that come from housing demands worldwide. These demands and the cost of land is influenced by several factors such as demography, economic growth government regulations and health facilities etc.

In the year 2007 and 2008 there was an economic collapse so there were several economic indicators that give the clue of disaster to follow, this is currently happening and the indicators suggest that the housing prices are getting high day by day. Almost 70% of the people who needs to buy houses are using the Internets to search so there is clear evidence that a relation between housing sales and housing prices exist.

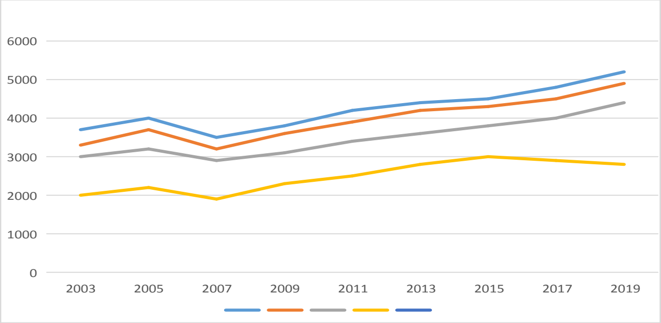
The real estate market is rapidly evolving. According to a published report by MSCI, Inc. it is estimated that the size of the professionally managed real estate investment market to be around $8.5 trillion next year, a total increase of $1.1 trillion than the previous year.

However, when we look at the market from a global perspective it turns out to be little simplistic. Although the market at a global scale is very tightly correlated, there are many aspects influencing the behavior of markets at a local scale, such as political instability or the emergence of highly demanded “hotspots” that can shift rapidly. In addition, different market segments evolve at different paces, such as industrial areas.

An example of these differences can be observed in the graph below which illustrates the change in price of land(per sq metre) in few Indian cities over a period of few years.

Blue: Bangalore, Orange: Lucknow,

Grey: Mysore, Yellow: Azamgarh



Looking at the figure, we can see some common patterns in the evolution of Bangalore’s and Lucknow, s markets, which are the two capital cities in India: a growth in the early 2000s, followed by a fall after 2007 due to the global financial crisis, which lasted until 2009, a moment after which prices have started to recover almost reaching maximum values again in 2013.

Meanwhile, in Mysore, which is a well-known vacational place, the increase in the last two years is more pronounced than in Bangalore and Lucknow. On the other hand, in Azamgarh, which is occupied mostly by rural areas, prices increased slightly till 2015, but are steadily decreasing since then. The figure clearly shows how although some global patterns can be found in the evolution, each region still shows some specificities in the prices evolution.

Of course, we can see higher variability while looking at specific assets. Most important factors driving the value of a house are the size and the location, but there are many other variables that are often taken into account when determining its value: number of bedrooms, availability of public transport (buses, underground, etc.), quality of schools in the area, shopping opportunities, availability of a lift (in apartments located in higher floors), availability of gardens or parks, etc. However, the main force that ultimately determines the value of houses is demand.

2. RESEARCH METHODOLOGY

2.1 The dataset

The dataset that we have chosen presents a few key challenges. Firstly, there is significant volume of data. More specifically, the database is composed of 12,223,582 total instances, such that 8,557,058 instances belong to the training set, whereas the remaining 3,666,524 are found in the test set. The advertisements refer to 2,300,079 distinct properties. As such, any approach that makes use of the whole data should be memory efficient. Moreover, each instance comprises twenty-four features of five different data types: integer, alphanumeric, date, string, float, and image. Therefore, any suitable model must be able to deal with mixed data types. Lastly, the significant amount of missing values for some key attributes made it particularly hard for regression, and the competitors had to come up with strategies for replacing the many missing values.

2.2 Description of dataset

The real estate housing data is used in this and it is taken from the machine learning repository and the data is spread across 20000 rows and attributes. The description of the data set is given below.

Table -1: Variable table

S.No.	Variables	Integer Type
1.	Latitude	Real
2.	Longitude	Real
3.	Housing Median Age	Integer
4.	Total Rooms	Integer
5.	Population	Integer
6.	Households	Integer
7.	Special attribute Y	Value to be predicted

Totally we are making use of 6 predictor variables and the Y variable here will be the median house price which we will predict.

2.3 Predicting the sale price

Now that we have described dataset as shown above we can start our plan of attack i.e. how to predict the sale price for a given house?

Linear regression models assume that the relationship between a dependent continuous variable Y and one or more explanatory or independent variables X is linear. It's used to predict continuous range values (e.g. price, sales) instead of trying to classify them into categories (e.g. animal, human etc.). Linear regression models are further divided into two main types:

2.3.1 Simple linear regression

Simple linear regression uses a traditional slope-intercept form, where a and b are the coefficients that we try to "learn" and produce the most accurate predictions. X is the input data and Y is our prediction.

2.3.2 Multivariable linear regression

Little bit more complex, multi-variable linear equation looks like this, where w represents the coefficients or weights, our model will try to learn.

$$Y(x_1, x_2, x_3) = w_1x_1 + w_2x_2 + w_3x_3 + w_0$$

The variables x_1, x_2, x_3 represent the attributes or distinct pieces of information, we have about each observation.

2.3.3 Random forest

Random Forest (RF) is a ML algorithm based on multiple decision trees whose outcomes are merged, leading to gains in performance and stability while still being faster than other ensemble methods, such as Adaboost since this method is robust to noise in the data, as well as able to tackle high bias and high variance. RF works in two steps: first, for every tree, we obtain a sequence of instances, sampled randomly with replacement from the training set. Each sequence of instances corresponds to a random vector Θk characterizing a particular tree. As the sequences will be slightly different from one another, so will the decision trees constructed from them. The prediction of the k-th tree for a given input x is given by:

$$h_k(x) = h(x, \Theta_k), \forall k \in \{1, 2, \ldots, K\}$$

where K is the number of trees. Each split of a tree uses a random selection of features to further avoid correlation. There are many ways to split a node S into two subsets. Assume a threshold c is chosen for the selected feature, splitting S into S1, S2 according to each feature value vi . For a regression task, one can use a c that minimizes the difference in the sum of squared errors:

$$SSE = \left(\sum_{i \in s_1} (v_i - \frac{1}{|s_1|} \sum_{i \in s_1} v_i)^2 + \sum_{i:i \in s_2} (v_i - \frac{1}{|s_2|} \sum_{i \in s_2} v_i)^2 \right)$$

The prediction of any subtree can be obtained as the mean (or median) output of instances that follow the same decision rules. The final prediction is simply an average of each tree's output:

$$h(x) = \frac{1}{K}\sum_{i=1}^{K} h_k(x)$$

3. CONCLUSIONS

The real estate market constitutes a good setting for investing, due to the many aspects governing the prices of real estate assets and the variances that can be found when looking at local markets. In this paper, we have explored the application of diverse machine learning techniques with the objective of identifying real estate opportunities for buyers and sellers. In particular, we have focused first on the problem of predicting the price of a real estate asset whose features are known, and have modeled it as a regression problem.

Predicting housing prices from online advertisements in is a task which requires insight into the data combined with powerful ML algorithms. In this work, we applied two different methods for this task, and combined them into a final prediction.

The Enriched RF works well with numeric features, as it can derive rules not only depending on the value of an attribute but also on its presence or absence. However, it cannot handle raw image or text data.

As future work, feature selection algorithms can also be employed to leverage the training speed for the models. Another technique which we aim to apply on this dataset is weak supervised learning with pseudo-labeling to increase the number of training data instances for deep learning.

REFERENCES

[1] Real Estate Price Prediction Using Machine Learning M.Sc Research Project Data Analytics

[2] Housing Prices Prediction with a Deep Learning and Random Forest Ensemble Bruno Klaus de Aquino Afonso, Luckeciano Carvalho Melo, Willian Dihanster Gomes de Oliveira1, Samuel Bruno da Silva Sousa, Lilian Berton

[3] Identifying Real Estate Opportunities Using Machine Learning Alejandro Baldominos, Iván Blanco, Antonio José Moreno, Rubén Iturrarte, Óscar Bernárdez and Carlos Afonso.

[4] Predicting House Prices with Linear Regression, Venelin Velkow.