Analysis of Western New York Real Estate Market

Swapnil Nitin Kamat  
Advisor: Dr. Carol J. Romanowski  
Department of Computer Science  
Golisano College of Computing and Information Sciences  
Rochester Institute of Technology  
Rochester, NY 14586  
snk6855@rit.edu

Abstract—Unlike several financial markets, transactions in real estate market involve huge amount of money and are usually long term investments. Owing to which, it is important for the stakeholders to have a sound understanding of the performance of real estate market at least over past few years, to minimize the risk of financial loss. This paper attempts to analyze different factors such as number of rooms, area in square feet, location, year built and previous price estimates of the non-commercial properties, and models a multiple linear regression to compute future price estimates based on the price variations observed in the locality corresponding to the properties. The dataset used for this analysis consists of non-commercial real estate properties within Western New York, especially the areas surrounding Buffalo metropolitan region.

Index Terms—Real estate, Multiple linear regression, Western New York

I. INTRODUCTION

The real estate market involves a vital phase of decision-making depending on the roles of stakeholders in the real estate domain. It is crucial for the Real estate investors to decide which geographic areas to target for investing in real estate properties to reap maximum returns in the future. The current owners of properties may be interested in knowing their house values an year ahead of time, to decide if they may want to sell or rent their property in the near future.

The real estate market of the United States has observed an uncertainty in property price variations in the past 5 years. Some geographic areas are observing a sudden steep rise in the property prices whereas some areas are still observing a steady trend. This uncertainty adds a risk factor to the decision making process, as the transactions of properties in real estate market involves huge amount of money. Multiple factors could influence the performance of real estate market resulting in variance of property prices. These factors can mainly be categorized as physical or environmental factors. Physical factors consist of physical aspects of the real estate property such as geographic location, land area in square feet, number of rooms, construction age and any recent improvements to the property. Whereas, environmental factors consist of proximity of resources such as schools, hospitals, marketplaces, downtown areas, public transportation or existence of any comparable properties in the neighborhood. Other factors such as increase in population density and economic indicators also contribute towards the price variance. To reduce the uncertainty in variation and to predict prices more accurately, all the aforementioned factors need to be studied and analyzed from the data captured over the past few years.

Since more than a couple of decades, property prices within Western New York region have been consistently low. According to the statistics provided by Zillow[1], a leading marketplace for real estate properties, areas adjacent to the Buffalo Metropolitan region are observing a major improvement in the local real estate market. These areas are gaining attention from real estate developers and investors, making the market even more competitive. This transition of the areas from being depressed with respect to property value to being on the verge of catching up with the high performing markets in the United States, is the reason behind selecting these areas for analytical purposes of this paper.

This paper discusses few approaches modeled in the past to predict property prices in the following background section. The methodology section briefly describes the overall architecture of the proposed solution and gives details about each component comprising the architecture. The results section discusses the outcome of experiments carried out in this paper. Finally, the conclusion and future work sections summarizes the results and provides suggestions that can be implemented as a part of future work.

II. BACKGROUND

In order to predict property prices, various approaches have been implemented in the recent past. Two of the most relevant techniques are discussed in this section.

[2] discusses an approach to predict current price of rental properties considering all the nearby landmarks such as schools, hospitals, tourist attractions, museums, theaters and other public places in the locality. The approach consists of two phases. In the first phase, clustering is performed over the rental properties which also overlaps the clusters of landmarks in the same locality. This clustering technique is based on three assumptions. The assumptions made are house price is higher if any landmark is closer to that house, the house price is directly proportional to the popularity of the nearest landmark and lastly, the same kind of facilities provided within community does not hike the house prices. The distance between landmarks and properties are calculated based on the Euler’s formula. To make prediction of rental
prices more specific, authors have implemented a multi-scale clustering technique which decides optimum size of the cluster to predict a reasonable price. In the second phase, a vector is calculated for each property, which encompasses features representing its distance from the nearest landmark and the landmark’s popularity. The house prices are then described by the multivariable linear regression model which takes these vectors as an input. At any given point in time, house prices can vary based on certain factors which cannot be foreseen. To account these unforeseen factors, authors have added a noise factor which would minimize the gap between the reasonable price and predicted price of the property. Although the approach discussed predicts a reasonable house price, the algorithm implemented only considers the present information about the landmarks and their popularity and does not consider the semantic attributes of the house itself. The approach also lacks the ability to predict the house prices in the near future.

[3] discusses an approach to forecast the performance of real estate in the near future using Artificial Neural Network(ANN). The author has selected time, average interest rate, change in sales, median house value, average number of days for the listing of house, inventory volume and inventory months supply to be the input variables for the ANN. The outcome of the ANN is the ratio of property’s selling and asking prices. The ANN is configured to have multiple hidden layers with varying number of hidden nodes for each of the layers. It uses feed-forward of input and back-propogation of error to train itself over the training dataset. Author uses different sizes of validation and test datasets with consisting of different instances to estimate the performance of the model. Several experiments are performed over the collected dataset by changing the number of hidden layers and nodes of ANN. Comparison of results prove that an ANN configuration with 1 hidden layer and 4 nodes performs well with error percentage within the range +2% to -2% for test instances. Although the model has a good training curve, it however does not guarantee good results when forecasting property prices more than 6 months ahead in the future.

III. METHODOLOGY

The approach used in the paper starts with the data collection phase. Once the data has been collected, it is analyzed to find correlations between attributes and then pre-processed before feeding it to the prediction model. The prediction model is implemented using multiple linear regression which takes several attributes from the dataset as input and attempts to compute an estimate for the price of the property. The approach can be broken down into various phases and each of these phases are discussed in more details in the following subsections.

A. Architecture overview

Architecture of the approach discussed in this paper consists of multiple components as shown in the Figure 1. The first component in the pipeline is data collection. The primary task of data collection component is reaching out the identified data sources and compiling the dataset by extracting the required attributes of the non-commercial real estate properties. However, the dataset obtained after the data collection phase can be termed as raw data. The raw data is then fed to the data pre-processing component. A series of data cleaning and preparation techniques are implemented to pre-process the raw data and transform it into well formatted data ensuring consistency and integrity.

Once the data has been cleaned, the next step is to partition the data into training and test datasets. The training dataset is then fed to the initial learning model component, which implements the multiple regression. Training of this initial learning model is done with k-fold cross validation. Finally the test dataset is used to predict the price estimates and compute accuracy of the final model. Following subsections discuss the actual implementation and working of multiple linear regression in detail.

B. Data Collection

Gathering relevant data is crucial to begin with the right analysis. Many data sources were searched over the web for property level details. The region selected for the analysis in this paper was the neighborhood of Buffalo metropolitan region. No pre-compiled datasets were found on-line consisting
of property details in the geographic areas of our interest. Hence, an approach was decided to crawl websites hosting real estate property listings. One such source identified was Zillow[1], is a leading real estate marketplace and lists over 110 million houses across the United States of America. Zillow hosts different categories of non-commercial properties that are listed for sale, rent or ones that are recently sold out.

The data collection phase was carried out using the following two approaches.

First approach was to implement a web crawler using Beautiful Soup and Selenium in Python. Beautiful Soup is a library which provides an object-oriented functionality to extract the required content from the markup of web pages. Whereas, Selenium is a library that provides functionality to automate the user generated events in a web browser. The crawler is featured to visit the homepage of Zillow and generate property listings for pre-defined zip codes corresponding to Western New York region. These listings represent properties which are available for sale, rent or have been recently sold out. The details page for each of these properties is visited by the crawler sequentially. All the information corresponding to the attributes listed in Figure 2 is extracted and added to the comma separated values file which is used as a dataset. This approach of collecting data consumes approximately 50 seconds to add a single instance to the dataset. The runtime of this technique is high as Selenium simulates the browser clicks and property details page is rendered for every property. An optimization was required to speed up the data collection phase and due to which the second approach was modeled.

Second approach is similar to the first approach but it eliminates Selenium. The rendering of each property details page using Selenium is avoided and replaced with API calls. Zillow provides an API which can be used to fetch details of properties specified by zip code and street address. However, the street address required for these API calls needs to be the exact street address of the property for which details are being requested. Thus, the web crawler first gathers street addresses by using Beautiful Soup library. Each of these addresses is then used to construct a query string for API call. An XML is sent back as a response to each of the API calls. XML response contains all the details for the property matching street address and zip code. A typical XML response is as shown in the Figure 3. The details included within the XML tags are extracted as required and added to the dataset.

The resulting dataset for the second approach includes non-commercial property details for over 7,204 properties across 25 zip codes in the Western New York region, surrounding Buffalo metropolitan area. Figure 4 visualizes the count of different types of properties present within the dataset. This dataset however does not include monthly or yearly price estimates for the properties. As this paper aims to build a prediction model for the property prices we need some previous price estimates for individual properties to train our prediction model. Hence, we add a derived attribute to the dataset to hold monthly price estimates for each of the properties. Zillow provides data consisting of monthly average for property prices grouped by zip code. For example, the average price estimate for zip code 14141 from August 2010 to November 2017 is as shown in Figure 5. We use this data to compute monthly price estimates since August 2010 till January 2018, which accounts for 90 months, for each of the 7,204 properties in our dataset. The final dataset obtained consists of 648,360 instances.
C. Analyzing Attributes and Finding Correlations

In this step, we analyze all the attributes in our dataset by calculating descriptive statistics and plotting these attributes against the attribute price estimate. For price predictions to be accurate, we need to find relationships between the price estimate and other attributes and feed only the significant attributes as input to the predictive model. Figure 6 shows how the average property prices increase as we move closer to the Buffalo metropolitan area. The attributes numOfBeds and numOfBaths have a directly proportional relation with the price as observed in the Figure 7. Also, the attribute finishedSqFt shares a positive linear relationship with the price of property. Although these relationships are logically obvious, it is always better to ensure that the instances in dataset depict expected relationships before we move on to the prediction phase.

D. Data Pre-processing

The dataset obtained from Zillow’s API is considered to be raw data and needs to be pre-processed before proceeding with the analysis and prediction phase.

1) Estimating Missing Values: The raw data includes multiple missing values for various attributes. These missing values need to be substituted or imputed with appropriate values ensuring data consistency. Few instances in the dataset have very less information as compared to the missing values. Such instances are removed from the dataset as the estimation of missing values would be incorrect for certain attributes if the other attributes for the corresponding instances are unknown. The attribute finishedSqFt shows a strong linear correlation with lotSizeSqFt. As a result the missing values for either of these attributes were best estimated based on the other known value among the two. The missing values for number of bedrooms and bathrooms were estimated manually based on the useCode and the finishedSqFt attributes.

2) Data Type Conversion: The learning model implemented for price prediction works on multiple linear regression technique. We thus convert all the required attributes to numeric data type as regression techniques work on numeric values. The attribute useCode holds a set of categorical values which are replaced with numbers decided with a proper mapping. The date string is converted to ordinal date value which represents a unique integer for corresponding date.

E. Regression Modeling

As this paper aims at modeling an approach for the price prediction of properties, we choose supervised learning techniques as they are more suitable for predictive models. A brief description on the working of multiple linear regression is attempted below. Multiple linear regression, as the name suggests, finds a linear relationship between variables. The input variables are
termed as independent or predictor variables and the output variable is termed as the dependent or response variable. Unlike simple linear regression which uses a single independent variable to predict the response variable, multiple linear regression uses two or more independent variables as inputs. The relationship between variables are modeled as a straight line termed as regression line, that best explains how the response variable changes with the change in multiple independent variables. A multiple regression model can be expressed as -

\[ y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon \]

where \( \beta_0 \) is the intercept, \( \beta_1 \beta_2 \ldots \beta_n \) are regression coefficients, \( x_1 x_2 \ldots x_n \) are independent variables, \( y \) is the dependent variable and \( \epsilon \) is called as error. The regression coefficients explain the relationship between the corresponding predictor variable and the response variable. The significance of each of the predictor variables can be measured with p-value statistics. The p-value is a decimal number ranging between 0 and 1, where the value close to 0 indicates low significance as compared to value close to 1. A predictor variable with p-value less than 0.05 is said to be insignificant. Attributes in the dataset identified to have high significance are \( \text{finishedSqFt} \), \( \text{numBeds} \), \( \text{numBaths} \) and \( \text{ZipCodeAveragePrice} \) over the less significant attributes \( \text{zipCode} \) and \( \text{yearBuilt} \). The results of prediction by this regression model are discussed in the following section.

IV. RESULTS

The test dataset used for computing prediction accuracy consists of 86,448 instances with property prices estimated for the year 2017. The accuracy of prediction was measured with two different considerations as presented in Figure 8. First, the model was configured to predict the exact prices of properties without any confidence interval. With this configuration any deviation of predicted prices with the corresponding actual prices were accounted as errors with the prediction. Out of total instances in the test dataset only 39,012 instances were able to be predicted correctly by the model. This resulted in accuracy of 45.12% which was lower than expected. However, it is not ideal to expect exact predictions for a continuous response variable from a prediction model, as even in the real world property prices observe some unexpected or unaccounted variations. Thus, an attempt at improving the accuracy was made by defining a confidence band of $1000. The instances for which the predicted property prices differ from the actual price by less than or equal to $1000 are said to be correctly predicted instances. By setting the confidence band, additional 28,684 more instances could be categorized as correctly predicted instances. This resulted in an overall accuracy of 78.6% which is an improvement of 33.48% over the previous configuration.

To analyze the variance in property prices with respect to the non-physical property aspects, the census data involving population estimates since 2010 till 2016 was obtained from US Census Bureau[4] for one of the counties in Western New York region, Cattaraugus County. The yearly average estimates of property prices were plotted against yearly population estimates for the same duration. The resulting graph is shown in Figure 9. From the graph we can interpret that as the population decreases the average house prices tend to increase. The article[5] explains the decline in population as the people moving-in into these areas are less as compared to the number of people moving out. However, the percentage of population loss is low than the previous years and thus the population is expected to remain fairly stable or even increase after few years in the future. Additionally, the article[5] mentions that influx of international immigrants in the counties surrounding Buffalo metropolitan area which partially explains the growth of property prices. Also, the former residents who own houses in more expensive markets are starting to return which can be a potential reason for price hike. The relationship between population estimate and average property prices can be analyzed as future work for significance of population estimate, if known, in the prediction of property prices.

V. CONCLUSION

Overall, it was interesting to gather the non-commercial real estate property details within Western New York region followed by implementation of prediction model for property prices. Following conclusions can be drawn from the results obtained -

- Multiple linear regression model, on this particular dataset, shows an improvement in the prediction accuracy when a confidence band is considered. The predicted property prices differ from actual prices by about 10% on an average.

<table>
<thead>
<tr>
<th>Consideration</th>
<th>Correctly predicted instances</th>
<th>Incorrectly predicted instances</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact price prediction</td>
<td>19022</td>
<td>47436</td>
<td>45.12</td>
</tr>
<tr>
<td>Price prediction with confidence band($\leq$ 1000)</td>
<td>67949</td>
<td>18499</td>
<td>78.6</td>
</tr>
</tbody>
</table>
Adding more independent variables to the multiple regression equation does not guarantee a better accuracy and could rather lead to over-fitting of the data. If the independent variables added to the regression are less significant with respect to the dependent or response variable then the accuracy of the model might decrease.

A strong correlation among the independent variables can lead to a less accurate prediction model. When two independent variables posses a strong relationship, it means that the change in either of these variables can anticipate the relative change in the other. As the change in both these variables is relatively proportional, either of these variables can bring the same change in the response variable. Thus we can avoid one of these variables and reduce the dimensionality of regression model without compromising on the overall accuracy of the model.

VI. Future Work

The approach used in this paper model tries to find a linear relationship between the input and response variables by implementing a linear regression. However, the variables might not necessarily possesses a linear relationship. Hence, different non-linear regression techniques can be implemented on the same dataset to compare the accuracy for each of these models.

The predictor variables used for multiple linear regression in this paper correspond to the physical factors of non-commercial real estate properties. Like the census data used in this paper to find correlation with the variance of property prices, we can find more such non-physical factors that show some relationship with the response variable. Inclusion of such factors in the regression equation can increase the accuracy of model and can be done as a part of future work.

ACKNOWLEDGMENT

I would like to thank my advisor Dr. Carol J. Romanowski for her valuable guidance and support throughout this project. I would also like to thank my colloquium advisor Professor Joe Geigel for tracking my progress and providing valuable feedback on my milestones.

REFERENCES