**NYC Housing Price Prediction**

Aanchal Agarwal, Abhiram Yenugadhati, Praneetha Moturi, Priyank Jagad

*San Jose State University*

*May 2021*

**Abstract – *Price prediction of properties is one of the most important applications in real estate as the housing market goes through a lot of up-and-downs due to volatile economic cycles. The aim of this project is to predict the house listing’s prices in the United States, specifically the state of New York City as that is where housing prices are most volatile. In this paper, data about the sale of properties was retrieved from the NYC finance department. The particular features of the nearby neighborhoods were used and models including regression analysis were implemented. Following which model optimization techniques such as model tuning, feature selection etc. were performed. The key problem was to improve the current models, which was addressed using techniques such as cross validation and hyperparameter tuning. The most optimum model was Cat Boost which showed a high R2 score (), a high recall () and a high precision () on the hyper parameter tuned data. Lastly, the effects of Covid-19 on the housing market was discussed.***

**Keywords:** Data Cleaning and Preprocessing, Exploratory Data Analysis, Regression Analysis, Feature Importance of the Best Model.

**I. INTRODUCTION**

The housing market is growing significantly and as such, a lot of real estate companies depend on the housing market for their bread and butter such as Real Estate Investment Trust. Such businesses also invest in the apartments and houses in New York state and try to regulate their internal pricing models. In New York City, the housing market including the property prices is quite buoyant due to a lot of macroeconomic reasons. However, it is pertinent to note that the inherent characteristics of the house also is a contributing factor to its pricing. Hence, for purchasers and house investors, having knowledge about the driving factors that influence the pricing of a house in United Statesis quite helpful in making wise purchasing decisions.

In this study to predict the housing pricing in the state of New York City, the data is treated with hyper parameter tuning and cross validation and then different models are implemented on the cleaned data. The preprocessed dataset is split

into train and validation set to evaluate our model’s performance in unseen data.

The regression analysis models under the umbrella of supervised learning methods that are implemented in this study are:

1. Linear Regression
2. Lasso Regression
3. Ridge Regression
4. Elastic Net
5. XGBoost Regression
6. Light Gradient Boosting Machine
7. K-Nearest Neighbors Regression
8. Decision Tree
9. Random Forest Regression
10. Cat Boost Regression

The optimization of our chosen model is done to the best accuracy with feature selection, model tuning and other techniques. The performance evaluation metrics that are used are R2 score, recall, precision and F1 scores to evaluate the best model with the best parameters. Meanwhile, the resulting model will create potential benefits in multiple areas beyond basic price prediction, such as:

1. providing data-proven insights for individual house buyers and sellers;
2. enhancing a balanced leverage between the buyers and sellers;
3. understanding the housing market in general for economists, policy makers or interested stakeholders/ decision makers.

**A. Dataset**

The dataset has been collected from the NYC department of finance which contains information about the sale of properties in New York City over a two-month period for the years 2020 and 2021. The dataset contains about 167720 property sales information pieces. Amongst most attributes we will have the Block, neighborhood, borough, Lot, Address, Apartment Number, Zip code. Every data instance contains information about demographics (address, region code, neighborhood), building information (type, number of units, building land area), sale date etc. There are 21 features in total. In this project, we will not study the effect of time on sale price, hence the feature “SALE DATE” will not be used to predict the sale price (target variable) of NYC property.

|  |  |  |
| --- | --- | --- |
| **Feature Name** | **Description** | **One Instance** |
| Borough | A digit code for the borough the property is located in; in order these are Bronx (1), Brooklyn (2), Manhattan (3), Queens (4), and Staten Island (5). | 1 |
| Neighborhood | The specific neighborhood the property is located. Department of Finance assessors determine the neighborhood name in the course of valuing properties. | Alphabet City |
| Building Class Category | The type of property. | 01 ONE FAMILY DWELLINGS |
| Tax Class at Present | The tax code of the property before transaction, includes the following: 1,2, 1A, 1B, 1C, 1D, 2A, 2B, 2C and 4. For example, class 2 properties include rental buildings, condominiums and cooperatives | 2A |
| Block | The digital code that represents the region the property is located in, commonly used with Lot and Borough (BBL) | 390 |
| Lot | The digital code that represents the street the property is located in, commonly used with Block and Borough (BBL) | 61 |
| Easement | An easement is a legal loophole that grants an interested party the right to use another person’s property or land in a certain way despite not having any ownership interest. | Nan (No Records in the Dataset) |
| Building Class at Present | The building code of the property before transaction, which indicates the type of building. For example, B1 indicates ‘TWO FAMILY BRIC’ | A1 |
| Address | The address of the property | 189 EAST 77TH STREET |
| Apartment Number | The apartment number of the property | 556 |
| Zip Code | The zip code of the property | 10009 |
| Residential Units | The number of residential units the property has | 2 |
| Commercial Units | The number of commercial units the property has | 1 |
| Total Units | The sum of residential and commercial units the property has | 5 |
| Land Square Feet | The usable or assignable square footage within the property, also known as net square feet (NSF) | 987 |
| Gross Square Feet | The space occupied by the intradepartmental circulation and the walls and partitions within the property, includes the land square feet | 2183 |
| Year Built | The year the property was built | 1998 |
| Tax Class at Time Of Sale | The tax code of the property during the transaction. The code description is the same as ‘Tax Class at Present’ | 2A |
| Building Class at Time of Sale | The building code of the property during the transaction. The code description is the same as ‘Building Class at Present’ | A1 |
| Sale Price | The specific time when the property is sold. | 5/23/2021 |
| Sale Date | The target variable. The sale price of the property, recorded in Canadian dollars. We have later converted this into US dollars. | $100000 |

**B. Data Preprocessing**

Data Preprocessing is the approach of data mining where the raw data is converted into an efficient and usable format to retrieve meaningful information from it.

**A. Steps Involved in Data Preprocessing:**

**(i) Data Cleaning**

Data Cleaning deals with the aspect of numerous missing and useless elements in the raw data. It deals with removing the noisy data and null values. In this project, the duplicates were first removed and the unique values were checked for each column. The data was then transformed wherein each column was converted to its respective data type, for instance, the land square feet to numerical. The “SALE PRICE” was also converted into US dollars. The EASE-MENT column was dropped initially since there was no information in it. Additionally, as the effect of time on the sale price is not considered in this project, the SALE DATE column was dropped as well.

The missing values were also dropped. After which isnull().sum() method was used to check whether the missing values were effectively dropped. It was observed that since the number returned was zero, the missing values were dropped efficiently.

**(ii) Outlier Detection**

Outliers are described as extreme values which deviate from the otherwise normal observations on data. They may indicate experimental errors, variance in the measurement, or a novelty. Hence, outliers are observations that differ from the overall patterns in the data.

**Methods used to treat Outliers:**

In this case, Z score is used to detect the outliers in the columns of “LAND SQUARE FEET”, “GROSS SQUARE FEET” and “SALE PRICE”. Z score is a significant measure that tells how much a number is above or below the mean of the dataset in terms of standard deviation. We set the threshold as 3. The count of outliers in “LAND SQUARE FEET”, “GROSS SQUARE FEET”, and “SALE PRICE” was detected as 79, 478, and 367 respectively. Hence, we dropped the outliers.

After the entire data cleaning process, the cleaned data contain 167783 rows and 19 columns.

**II. EXPLORATORY DATA ANALYSIS**

**A. Feature Scaling**

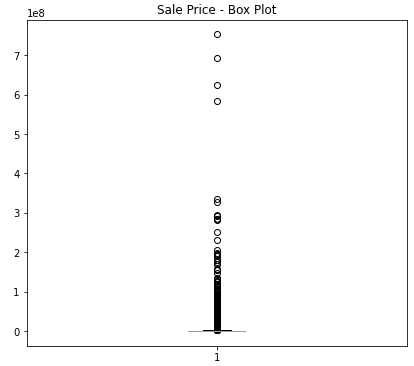
Feature Scaling transforms the data into a format that can be used and worked on in the mining process. Since we have used R squared scores as a performance metric, we have implemented normalization techniques to scale the data such as Min-Max Normalization of the numerical features in the range of -1 to 1.

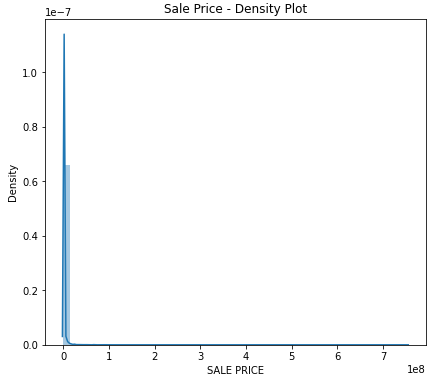
**C. Data Visualization**

Data visualization is the graphical depiction of data and information which makes it easier to comprehend and examine the patterns and trends in the data by the use of visual elements such as maps, graphs and charts.

**Target Variable (Sale Price)**

In the present case, our target variable will be sales price and the remaining features will help us to predict sales price for unseen data It is observed that the distribution of sale price from the raw price is significantly sparse. The mean, median and mode for each column in the dataframe was calculated. It is seen that a lot of sales occur with an absurdly small number: $0 most commonly (note that 40% of the sale price is $0). On the basis of the original data source, it is noted that the sales are in effect transfers of the deeds between parties: for instance, the transfer of ownership of the house from parents to children after the parents move out for retirement. To handle this situation, a reasonable range for the sale price is set up. The instances for which the sale price is less than $50000 (41% of the entire data) and greater than $12M (Notice that the $12M threshold helps eliminate the 0.85% special cases) will be removed since it will help eliminate the special cases. Following which, log transformation is performed since the numbers are huge.

****

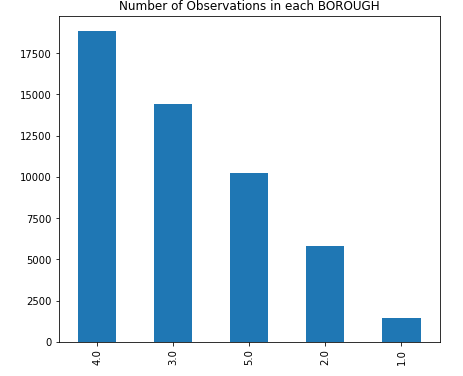
****

**Fig. 1.** Distribution of building sale prices before data cleaning and log(x) transformation

**D. Predictive Feature Analysis**

Within the scope of this study, it is noted that the features Borough, Neighborhood, Block, Lot, Address, Zip Code and Apartment Numbers are associated with the location of the properties. Since they are highly correlated with each other, after careful consideration, Borough was the only location feature that we kept. There are five Boroughs in our dataset:

* Bronx
* Brooklyn
* Manhattan
* Queens
* Staten Island



**Fig. 2.** Number of buildings in each Borough (1 = Bronx, 2 = Brooklyn, 3 = Manhattan, 4 = Queens, 5 = Staten Island)

***Observations***: It is observed that Queens has the most data instances whereas Bronx has the least.

Block represents the region of the property and Lot represents the street a property locates. Both Block and Lot are often used together with Borough (called a Borough-Block-Lot location system). Similarly, Apartment Numbers, Zip Code and Addresses each have 6670, 195 and 159351 unique values. The features discussed above except for Borough are very sparse and highly correlated with the Borough. Therefore, for the purpose of this project, only Borough is considered as the predictive feature.

The features of Building Class Category, Building Class at Present and Building Class at Time of Sale describe the types of property wherein the latter two are sparse and are mere subdivisions of the Building Class Category. For model simplicity, we kept Building Class Category only. By looking at Figure 3 in Appendix B, different types of family dwellings and apartments with elevators are the most frequent building types, meaning that most buildings are of residential uses. By looking at Figure 4, Appendix B, it seems that certain building classes have a larger range of prices (e.g. Rentals - Apartments with Elevators), or some higher average prices in general (e.g. Luxury Hotels).

**5. Goals of the Project**

**6. Technical Aspects**

**6.1 Enhanced Version of Working**

**8. Conclusion**