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 were 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 XG Boost which showed a high R^2 score on the hyperparameter tuned data. Lastly, the effects of Covid-19 on the housing market were 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 States is 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:

A. Linear Regression
B. Lasso Regression
C. Ridge Regression
D. Elastic Net
E. XGBoost Regression
F. Light Gradient Boosting Machine
G. K-Nearest Neighbors Regression
H. Decision Tree
I. Random Forest Regression
J. Ada Boost Regression
K. 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 R^2 score, MSE and RMSE 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	Represents the borough's digit code in which the property is located – These ae ordered as Bronx (1), Brooklyn (2), Manhattan (3), Queens (4), and Staten Island (5).	1
Neighborhood	The specific neighborhood the property is located in. The name of the neighborhood in the course of valuing the properties is determined by the Department of Finance assessors.	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	It is a legal loophole in which the interested party	Nan (No Records in the Dataset)

	can avail the right to use other person's property without any interest of ownership.	
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.

B. 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 visualization 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 data frame 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.

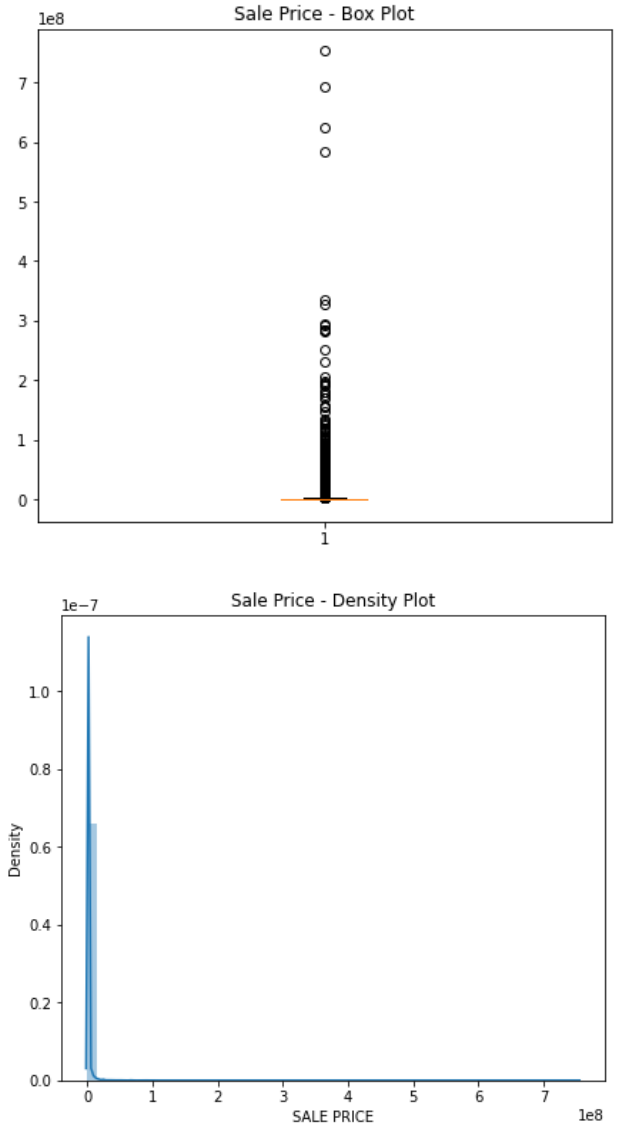

Fig. 1. Distribution of building sale prices before data cleaning and log(x) transformation

C. 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

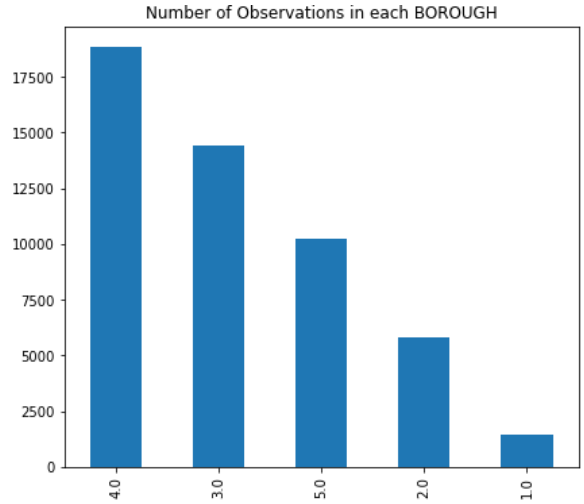

Fig. 2. Number of buildings in each Borough (1 = Bronx, 2 = Brooklyn, 3 = Manhattan, 4 = Queens, 5 = Staten Island)

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 basically subdivisions of the Building Class Category and are sparse. To keep the model simple, we have implemented the Building Class Category alone.

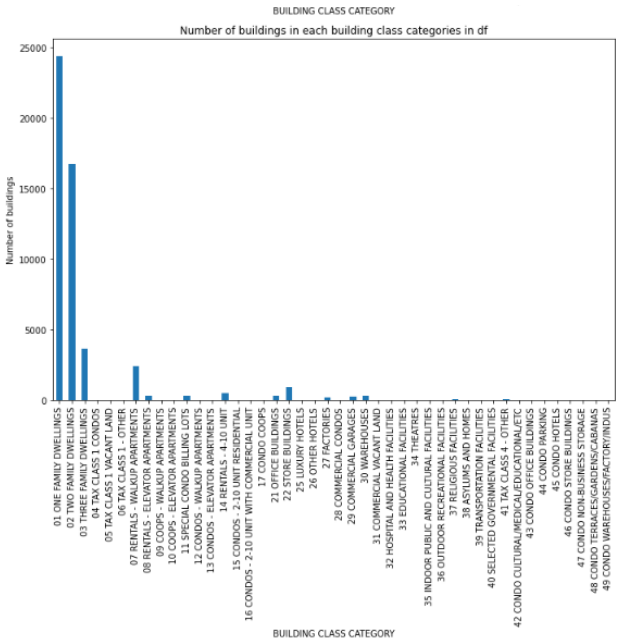

Fig.3. Number of buildings in each Building Class Category

Observations: From the above figure, we can observe that the most frequent building types are the different types of family dwellings such as one, two, and three-family dwellings and the rental apartments. This implies that most of the buildings are of residential nature and use.

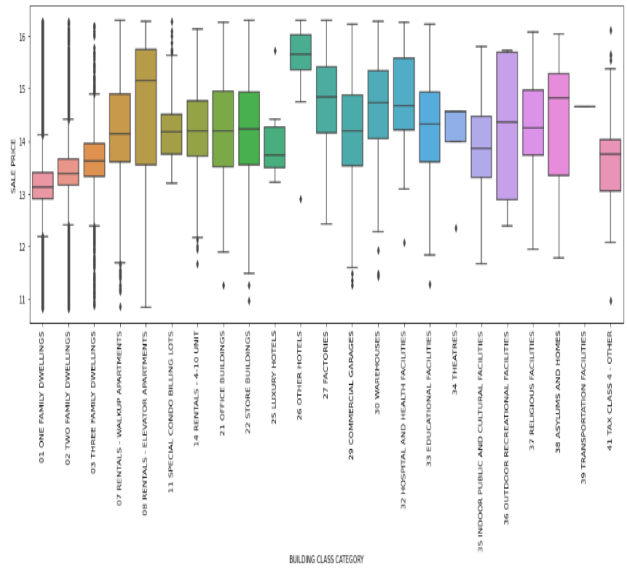

Fig.4. Boxplot of log(Sale Price) against Building Class Category.

Observations: From the above figure, it is observed that there are some interesting patterns present between the building types and their sale prices. It is noted that certain building classes comprise of a larger range of prices or some higher average prices such as Rentals – Elevator Apartments.

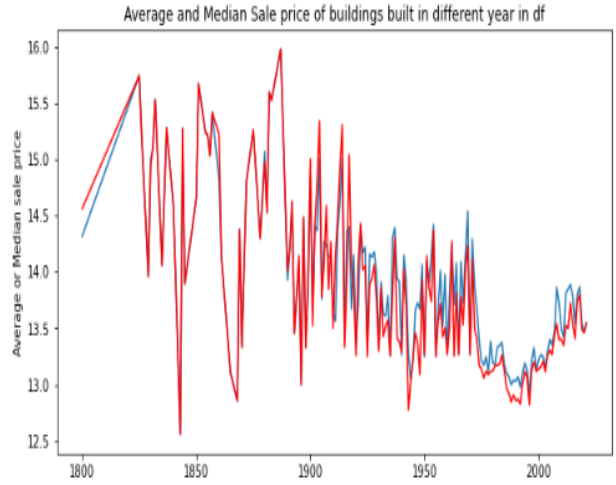

Fig.5. Mean log(1 + Sale Price) and Median log(Sale Price) against the year in which the building is built

Observations: It is observed that prior to 1900, the sale price of properties that were built were lower than the ones that were built after 1900. Hence, it can be said that Sale Price is dependent on the Year Built.

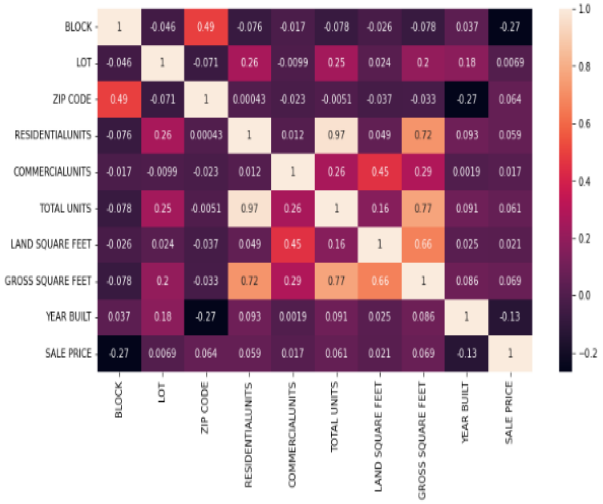

Fig.6. Correlations between numerical variables before Feature Engineering

Observations: From figure 6, it is observed that the correlation between Residential units and Sale Price, which is our target variable is linear and positive. The same pattern also exists in commercial units, which is the number of commercial units in a property, and the total units which is the sum of residential and commercial units.

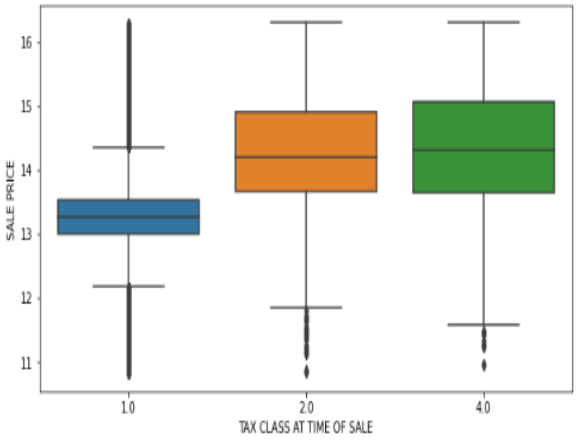

Fig.7. Boxplot of log(1 + Sale Price) against Tax class at time of sale

Observations: It is observed from the above that the box plot of the transformed sale price against the Tax Class at Time of Sale shows three unique tax classes at the time of sale.

- Tax class 1: More right-skewed with more number of high sale price outliers. It also has smaller IQR(Interquartile Range) with the lowest median sale price.
- Tax class 2: Fewer high sale price outliers. It also has a larger IQR and a higher median sale price.
- Tax class 4: No high sale price outliers. It has the largest IQR and the highest median sale price.

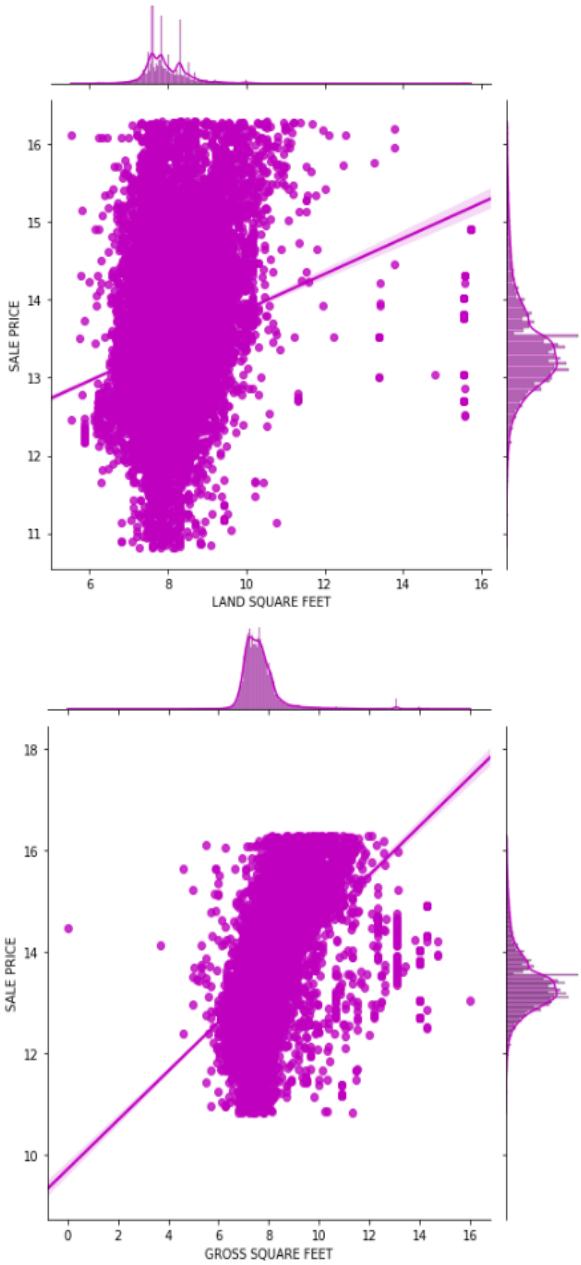

Fig.8. Relationship between property land square feet and sale price

Observations: It is observed that the land square feet and gross square feet share similar distribution. They correlate positively with the property sale price. Nonetheless, we also observed that when the land square feet are small, the sale price is high. The outliers are few yet possible explanations for such outliers could be that they belong to special building classes or they lie in a very good geographical location.

D. Feature Selection

Based on the correlation heatmap in figure 6 and the discussion in Exploratory Data Analysis, the following columns were dropped:

- Neighborhood
- Address
- Apartment numbers
- ZIP code
- Building class at Present
- Building class at Time of Sale
- Tax class at Present
- Sale Date.

Additionally, the Easement column was dropped as well since it only contains null values.

E. Feature engineering Classification:

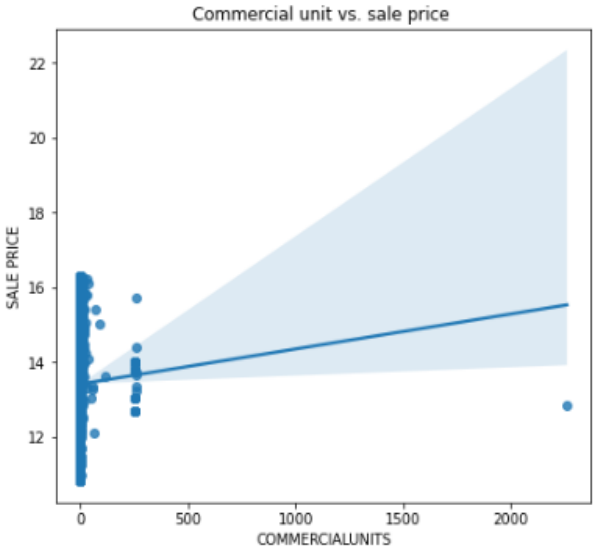

Fig.9(a). Scatter Plot of CommercialUnits vs. Sale Price

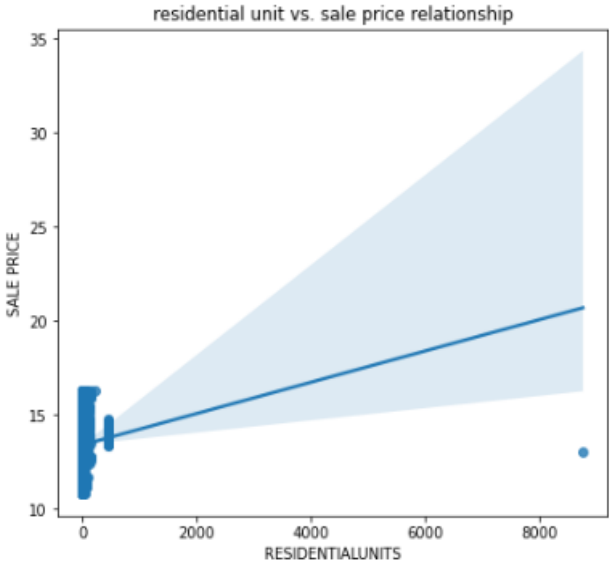

Fig.9(b). Scatter Plot of ResidentialUnits vs. Sale Price

Observations: It is observed that the pattern is opaque and there are a lot of 0s and 1s in each plot. Thus, we classify the CommercialUnits and ResidentialUnits into six groups.

UNIT TYPE	CRITERIA
A	Commercial Units > 10
B	0< Commercial Units <= 10
C	Commercial Units = 0 and Residential Units= 1
D	Commercial Units = 0 and 1 < Residential Units < 10
E	Commercial Units = 0 and Residential Units >= 10
F	Commercial Units = 0 and Residential Units

Table 1 - Grouping commercial units as a categorical variable

Therefore, we introduce a new variable named 'UNIT CATEGORY" which represents the pattern of COMMERCIALUNITS and RESIDENTIALUNITS.

F. Categorical features & One-hot encoding:

To build the models, we employ one-hot encoding in order to transform the features of

BOROUGH, BUILDING CLASS CATEGORY, TAX CLASS AT TIME OF SALE and UNIT CATEGORY. It is observed that after one-hot encoding, we have 50706 instances with 62 columns, which is a little sparse. Hence, after building the models, we will evaluate the performance using the metrics.

G. Numerical feature - Rescaling:

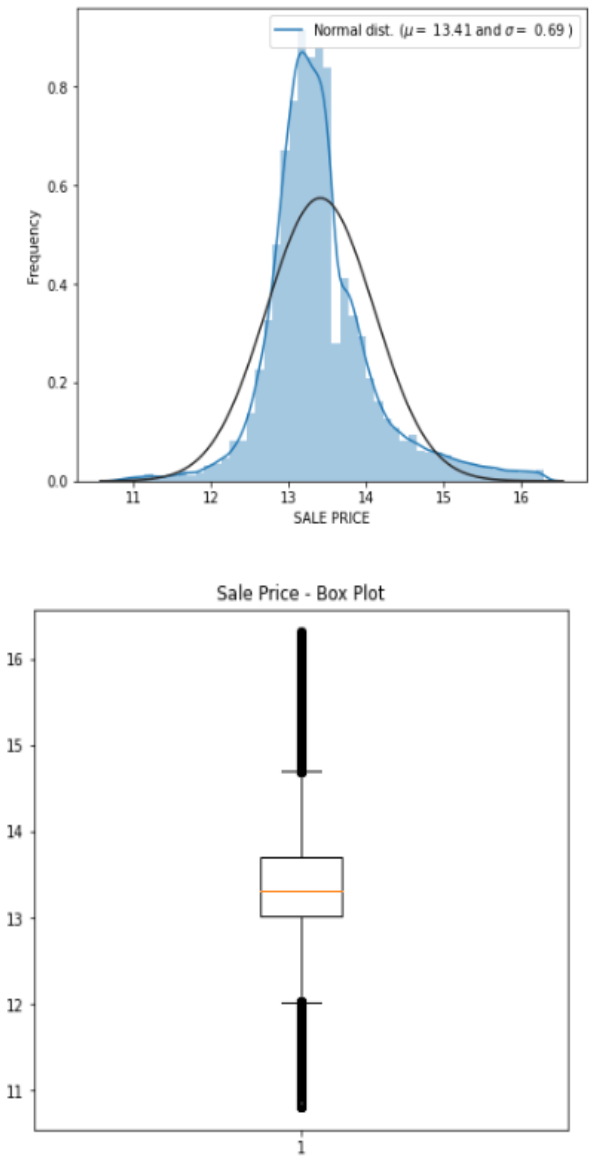

Fig.10. Distribution of Sale Prices after log transformation

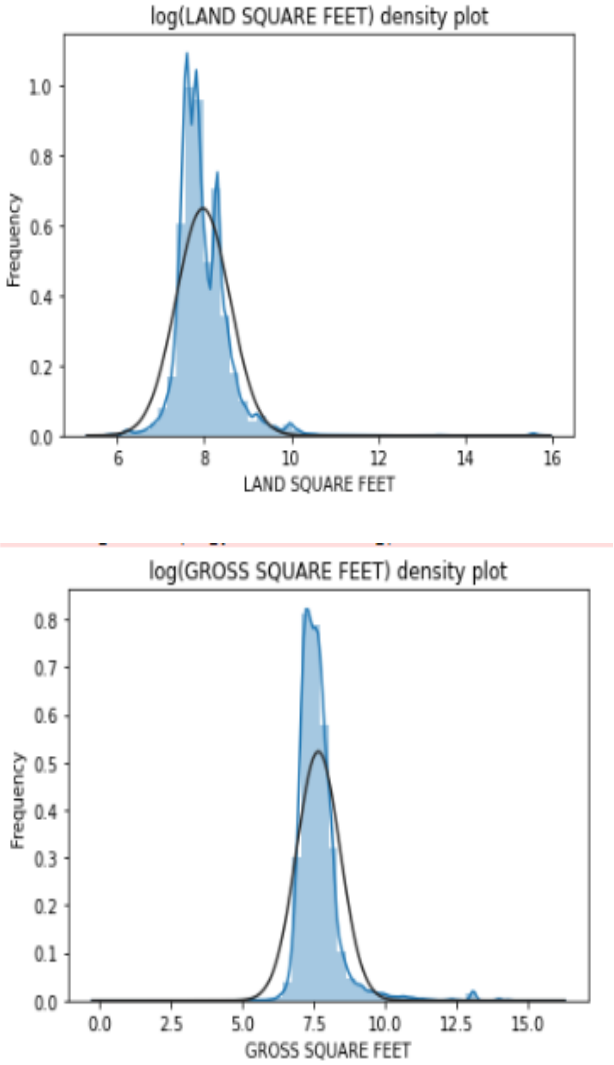

Fig.11. Distribution plots of land square feet and gross square feet after log transformation

Observations: Based on the above figures which show the density plots of Sale Price, Land Square Feet and Gross Square Feet, the distribution is sparsely allocated and heavily right-skewed. Therefore, log transformation is performed on these three features.

III. MODELS IMPLEMENTED

A. Linear Regression

When we fit the best fit line (usually straight) through the data, it is known as linear regression. Fitting lines to non-linear data will result in

different levels of overprediction and underprediction. In this project, we assume that the target variable sales price has an expected linear relationship between various variables. But to examine and understand the true structure of this data, it is recommended that a polynomial curve is fit into our data. One improvement would be a technique where new features are engineered into existing input variables functions (including logs, powers, and products of pairs of variables).

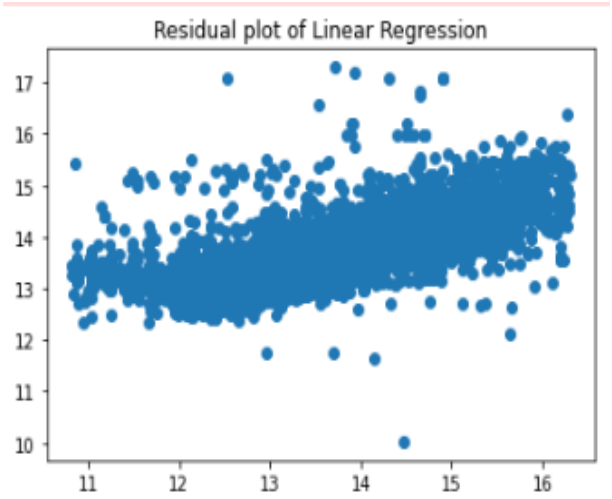

Fig.12. Residual Plot of Linear Regression

B. Lasso Regression

Lasso regression is a type of linear regression that uses shrinkage. Shrinkage is where data values are shrunk towards a central point, like the mean. The lasso procedure encourages simple, sparse models (i.e. models with fewer parameters). This regression is well-suited for models showing high levels of multicollinearity or when you want to automate certain parts of model selection, like variable selection/parameter elimination.

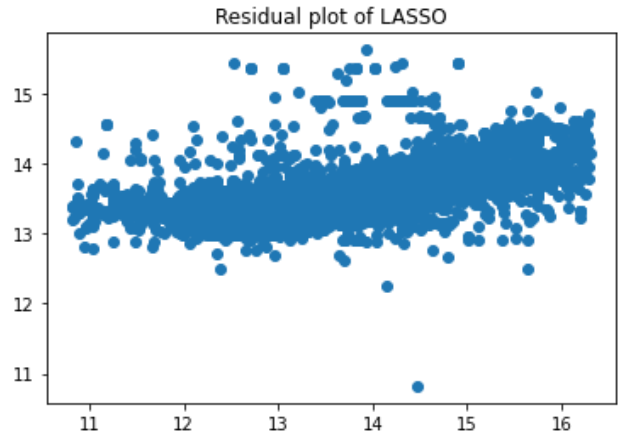

Fig.13. Residual Plot of Lasso Regression

C. Ridge Regression

When the number of predictor variables in a set exceeds the number of observations, or a data set contains multi-collinearity, Ridge regression is an approach to develop a parsimonious model.

Ridge regression employs a ridge estimator, which is a special type of shrinkage estimator. Shrinkage estimators generate new estimators that are closer to the "actual" population parameters in theory. The Ridge estimator is especially good at improving the least-squares estimate when multi-collinearity is present.

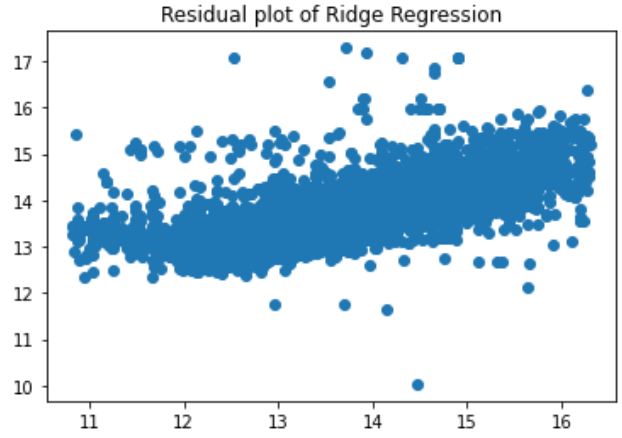

Fig.13. Residual Plot of Ridge Regression

D. Elastic Net

Elastic net linear regression uses the penalties from both the lasso and ridge techniques to regularize regression models. The technique combines both the Lasso and Ridge regression methods by learning from their shortcomings to improve the regularization of statistical models.

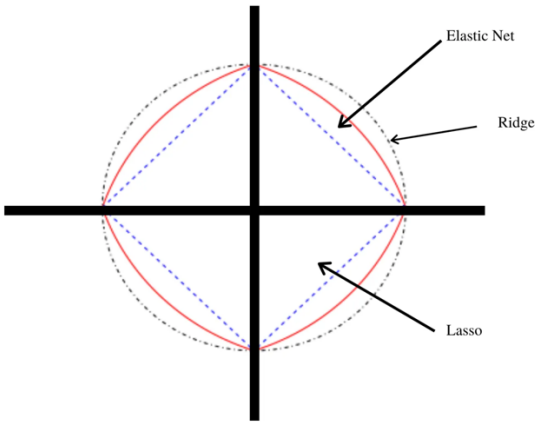

Fig.13. Comparison between Ridge, Elastic net and Ridge

The elastic net method overcomes lasso's constraints, such as when high-dimensional data requires only a few samples. The elastic net approach allows "n" variables to be included until saturation is reached. If the variables are highly connected groups, lasso will usually pick one from each group and ignore the others.

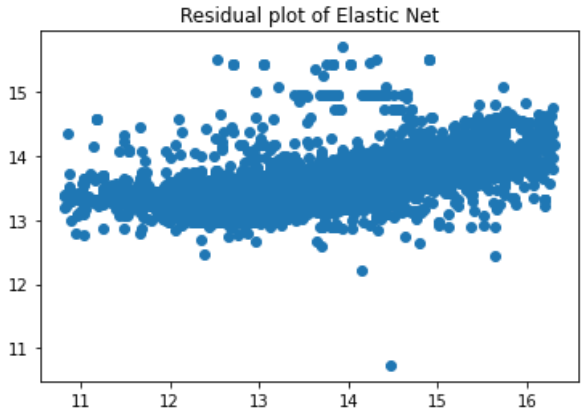

Fig.14. Residual Plot of Elastic Net

E. XGBoost Regression

XGBoost is a popular and efficient open-source implementation of the gradient boosted trees algorithm. Gradient boosting(minimize the loss when adding new models) is a supervised learning approach that combines the estimates of a set of smaller, weaker models to attempt to accurately predict a target variable. The weak learners in gradient boosting for regression are regression trees, and each regression tree transfers an input data point to one of its leaf containing a continuous score.

XGBoost combines a convex loss function (based on the difference between the anticipated and target outputs) with a penalty term for model complexity to minimize a regularized (L1 and L2) objective function (in other words, the regression tree functions).

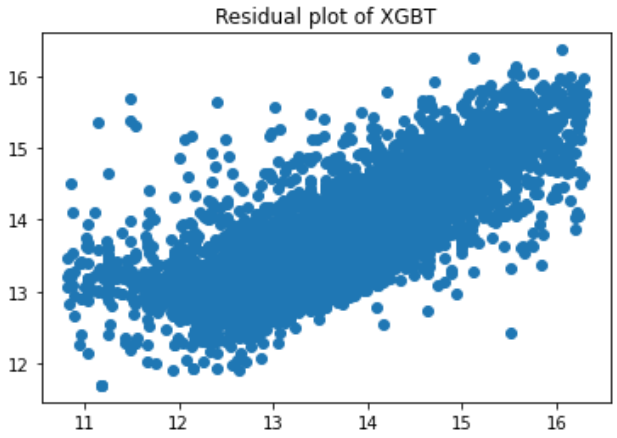

Fig.15. Residual Plot of XG Boost

F. Light GBM Regression

LightGBM improves on the gradient boosting technique by incorporating automatic feature selection and concentrating on boosting examples with larger gradients.

This can lead to a significant increase in training speed and enhanced prediction performance. As a result, when working with tabular data for

regression and classification predictive modeling tasks, LightGBM has become the de facto approach for machine learning contests.

As a result, along with Extreme Gradient Boosting, it shares some of the blame for the growing popularity and wider acceptance of gradient boosting methods in general (XGBoost).

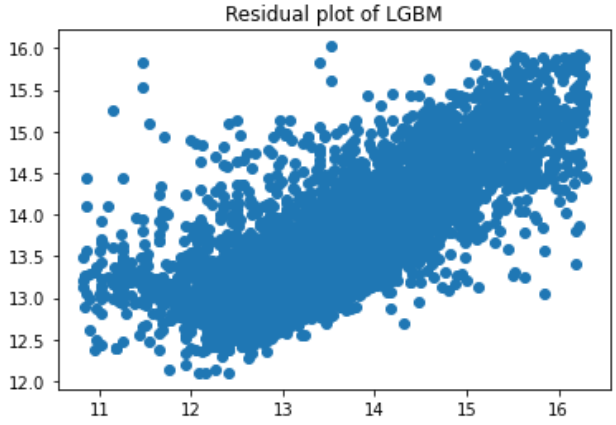

Fig.16. Residual Plot of Light GBM

F. K-Nearest Neighbors Regression

KNN regression is a non-parametric method that approximates the relationship between independent variables and continuous outcomes by averaging data in the same neighborhood in an understandable manner. The analyst must set the size of the neighborhood, or it can be decided using cross-validation (which we will see later) to find the size that minimizes the mean-squared error.

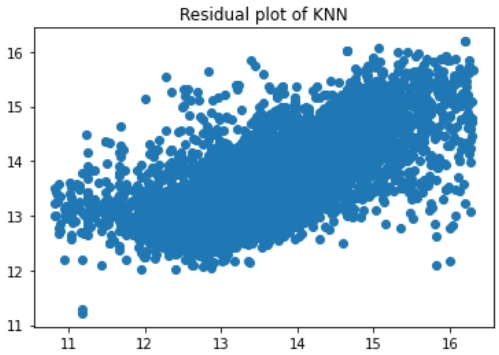

Fig.17. Residual Plot of K-Nearest Neighbors Regression

G. Decision Tree Regression

Decision tree builds regression models in the form of a tree structure. It incrementally breaks down a dataset into smaller and smaller sections while also developing an associated decision tree. A tree with decision nodes and leaf nodes is the result. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data.

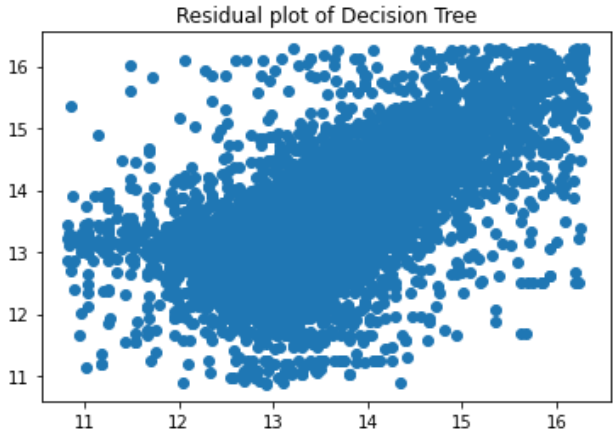

Fig.18. Residual Plot of Decision Tree Regression

G. Random Forest Regression

Random Forest Regression is a supervised learning algorithm that uses ensemble learning method for regression. Ensemble learning method is a technique that combines predictions from multiple machine learning algorithms to make a more accurate prediction than a single model. A Random Forest Regression model is powerful and accurate. It usually performs great on many problems, including features with non-linear relationships.

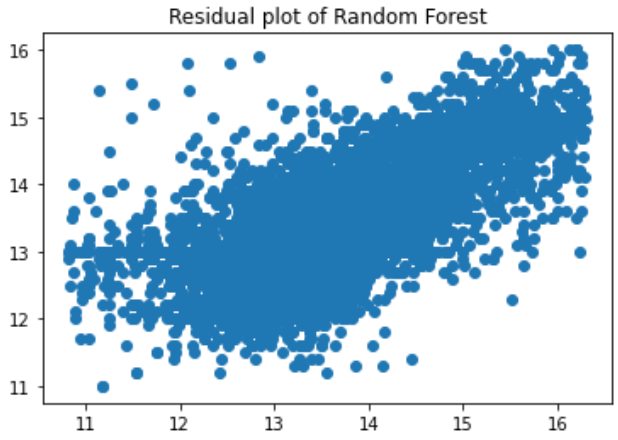

Fig.19. Residual Plot of Random Forest Regression

H. Cat Boost Regression

Cat Boost builds upon the theory of decision trees and gradient boosting. Because gradient boosting fits the decision trees sequentially, the fitted trees will learn from the mistakes of former trees and hence reduce the errors. This process of adding a new function to existing ones is continued until the selected loss function is no longer minimized. Cat Boost also offers an idiosyncratic way of handling categorical data, requiring a minimum of categorical feature transformation, opposed to most other machine learning algorithms, that cannot handle non-numeric values.

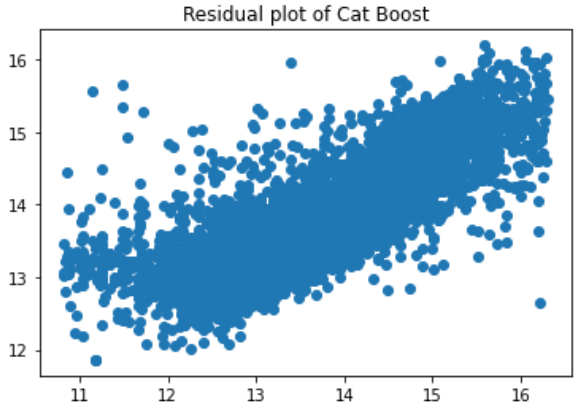

Fig.20. Residual Plot of Cat Boost

I. Ada Boost Regression

AdaBoost (Adaptive Boosting) is a very popular boosting technique that aims at combining multiple weak classifiers to build one strong classifier. AdaBoost uses an iterative approach to learn from the mistakes of weak classifiers and turn them into strong ones.

A single classifier may not be able to accurately predict the class of an object, but when we group multiple weak classifiers with each one progressively learning from the others' wrongly classified objects, we can build one such strong model. The classifier mentioned here could be any of your basic classifiers, from Decision Trees (often the default) to Logistic Regression, etc.

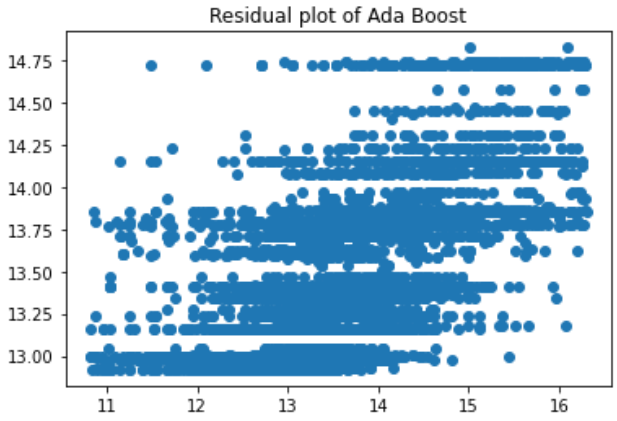

Fig.21. Residual Plot of Ada Boost

Model Results:

Model	**R^2**	**MSE**	**RMSE**
Linear Regression	0.4814	0.2547	0.5047
Lasso Regression	0.2909	0.3483	0.5902
Ridge Regression	0.4818	0.2545	0.5045
Elastic Net	0.2978	0.3450	0.5873

XGBoost Regression	0.6544	0.1697	0.4120
Light Gradient Boosting Machine	0.6290	0.1822	0.4269
K-Nearest Neighbors Regression	0.5172	0.2371	0.4870
Decision Tree	0.3138	0.3371	0.5806
Random Forest Regression	0.6298	0.1818	0.4264
Ada Boost regression	0.3496	0.3195	0.5652
Cat Boost Regression	0.6382	0.1777	0.4216

IV. COMPARISON

A. Performance Metrics

1. R-squared means – It represents the proportion of dependent variable variance which is described by the independent variable. It is a statistical measure.

2. MSE – It is the mean squared error and it indicates how close the data points are to a regression line.

3. RMSE – It is the root mean square error and indicates how far the data points are from the regression line.

B. Best model evaluation

Among all the models that were trained and implemented, XG Boost is the winner as it offers better performance on the data. The R squared means is 0.6544 which is by far the best among all the models that were implemented. The features that were better predicted for the highest importance in this model are:

- Gross Square Feet
- Land Square Feet
- Borough
- Year Built
- Age

Hence, the housing price prediction of the properties in New York City was performed.

IV. CONCLUSION

In conclusion, we observe that XG Boost outperforms all the models with the highest r R^2 score of 0.6544 and an MSE of 0.1697.
One of the challenges we faced was the data collection and lack of extensive and appropriate data. Hence, if allowed more time and appropriate data, we may incorporate the effect of COVID-19 on housing prices after 2020. Regarding the future scope of this project, it can be extended to multiple states and this could be further developed into an API for implementation and price estimation of Real estate applications. Hence, property price prediction models add a lot of value to the field of real estate.

V. REFERENCES

[1]. Dataset Link: https://www1.nyc.gov/site/finance/taxes/property-annualized-sales-update.page

[2]. Research Paper

C. R. Madhuri, G. Anuradha and M. V. Pujitha, "House Price Prediction Using Regression Techniques: A Comparative Study," 2019 International Conference on Smart Structures and Systems (ICSSS), 2019, pp. 1-5, doi: 10.1109/ICSSS.2019.8882834.