**OPIM – 5604 – Predictive Modeling**





**PREDICTING HOUSE PRICES IN BROOKLYN**

**Team-6**

**Alekhya Uppugandla**

**Arkaprava Kundu**

**Ashish Tele**

**Harshvardhan Agarwal**

**Manish Arora**

**Contents**

[Problem Statement/Objective 3](#_Toc512619543)

[Source 3](#_Toc512619544)

[Exploratory Data Analysis 3](#_Toc512619545)

[Characteristics that affect housing prices 4](#_Toc512619546)

[Data Modification 6](#_Toc512619547)

[Log Transformations 8](#_Toc512619548)

[Normalizing and Standardizing the data 10](#_Toc512619549)

[One-Hot Encoding: Categorical Data 11](#_Toc512619550)

[Modeling 12](#_Toc512619551)

[(a)Linear Regression: 12](#_Toc512619552)

[(b) Regularized regression: Lasso and Ridge: 13](#_Toc512619553)

[(c)Random Forest: 14](#_Toc512619554)

[(d)Boosted Tree (Xgboost): 15](#_Toc512619555)

[Learning Curve 17](#_Toc512619556)

[Recommendations 17](#_Toc512619557)

[1)Recommendations for customer who are investors: 17](#_Toc512619558)

[2)Recommendations for Builders: 18](#_Toc512619559)

# 

# Problem Statement/Objective

**Introduction**: New York City's shortage of affordable housing has reached a crisis point. Especially in Brooklyn, a demand supply gap has led to a continuous increase in house prices. The prices are determined by several factors, including size of the apartment, neighborhood, proximity to commercial hubs and other amenities etc. Brooklyn housing sales prices has seen a continuous increase from 2003 to 2017, and these variables have played a major role in determining the sales prices.

**Objective:** To get the approximate sale prices for houses in Brooklyn in the next few years by utilizing regression techniques and visualization

# SEMMA Approach to Predictive Modeling

Every modeling project must follow SEMMA approach i.e. once after the business goal is defined we need to follow the steps of SEMMA. In our project we made sure to follow these steps and draw insights from this accordingly.

# Source

The data has been taken from following link in Kaggle

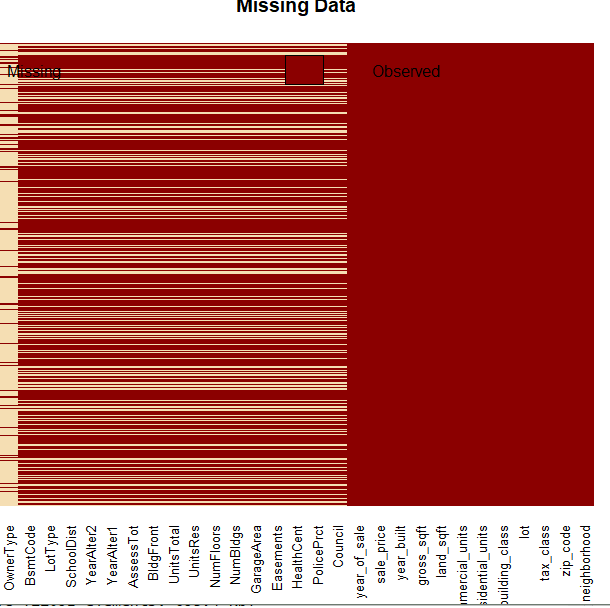
<https://www.kaggle.com/tianhwu/brooklynhomes2003to2017>

The primary dataset for the Housing Sales data has been taken from the NYC Department of Finance site. In addition, the data for other important housing variables has been obtained from NYC Department of City Planning.

# Exploratory Data Analysis

The initial data consisted of 390883 rows, and 109 variables along with the sales\_price variable. To remove redundant variables, an analysis of the variables affecting sales price was done and we brought down the number of independent variables to 27.

Exploration of the data was accomplished using data visualization techniques in R. Bar plots were used for categorical variables and scatterplots and qqplots proved to be useful to understand the continuous data relation with sale price.



**Removing missing values**

Due to the unstructured nature of the data, many of the data had missing values and were improperly arranged.

As we observe in the missing map graph below, 17 categorical variables had many missing values. It’s difficult to impute the values for these columns as all predictors are blank in that specific row. Hence, we dropped the observations (22.3% of the value)

|  |
| --- |
|  |
|  |
| |  | | --- | |  | |

# 

# Characteristics that affect housing prices

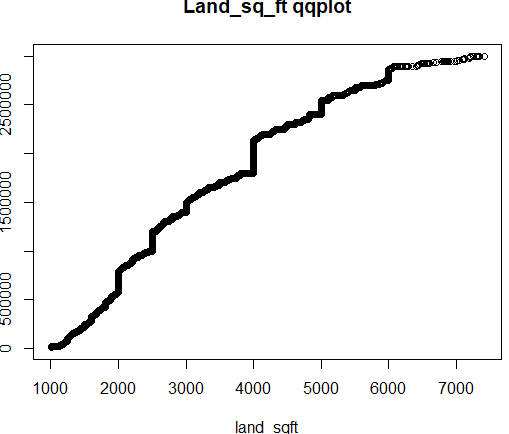
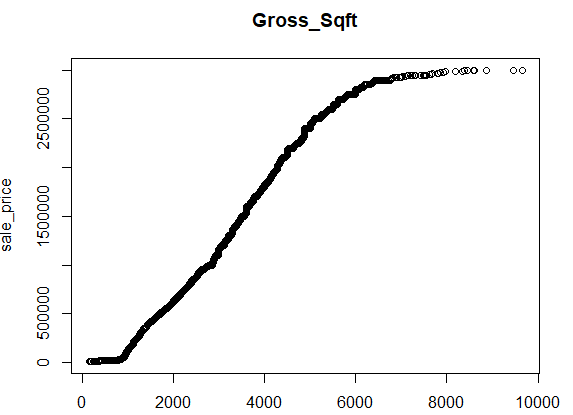
In terms of the housing characteristics that were used in our regression models, we wanted a well-rounded list of housing characteristics encompassed in these three broad categories: Property Characteristics, Community Characteristics and Proximity Characteristics. We wanted to include more than just structural variables in our hedonic regression models.

**Property Characteristics**.

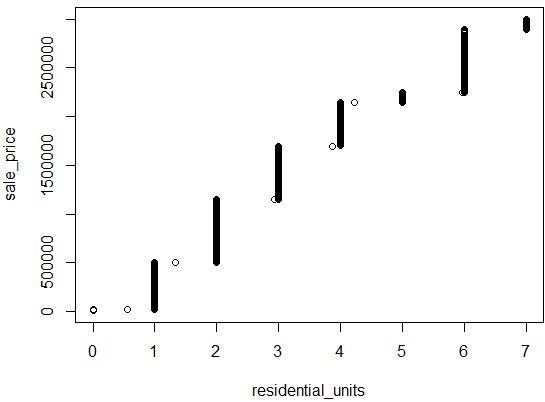
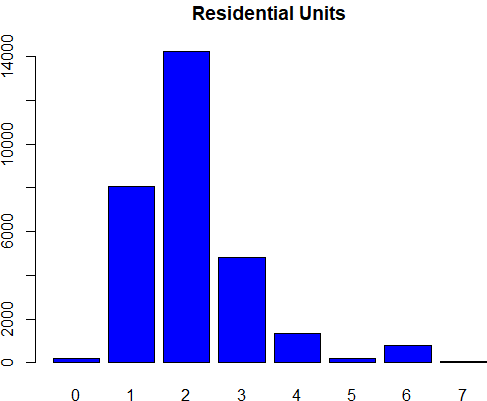
The housing prices are largely dependent on the house characteristics. In general, larger homes with more floors and area size tend to sell for higher prices. The number of bedrooms and bathrooms tend to increase the sale price, even after controlling other physical, locational and quality features. Area of the property is highly proportional to the sale price of the house. From our data, we plotted the different house characteristics against the sale\_price and came to some conclusions about their relation. Notably the housing characteristics variables in our data are the following

Landsqft, Gross Sqft, Bldg Front size, Lot Size, Residential Units, commercialunits, NumFloors, NumBuildings, UnitsRes, GarageArea, YearAlter.

After doing EDA of the variables we observed the following characteristics:

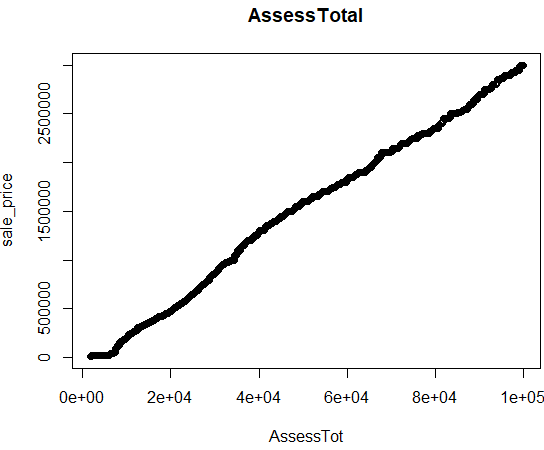
 

The housing price increases with the area of the land sqft, i.e. they are directly proportional, while the housing price in general increases with gross sqft increase till 8000 sqft, then it remains almost constant.



The number of residential units is mostly from 1 to 3, with very les units above 4. But from the right-hand graph, we observe that the sales\_price in general increases with the number of residential units

Most of the commercial units is 0, and very less units above 2. But from the right-hand graph, we observe that the sales\_price in general increases with the number of commercial units



The sale\_price shows a very good linear relation with the Assesses Total Price, i.e. the sale price is in line with Assessed Total price.

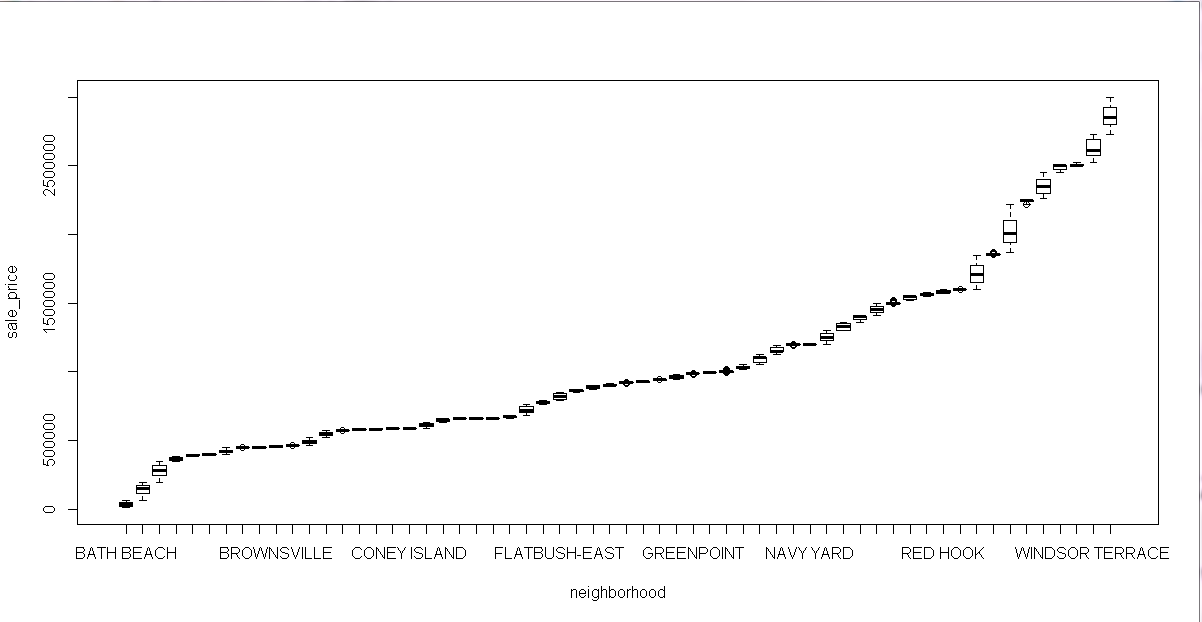
**Neighborhood Characteristics**

In addition to the housing characteristics, surroundings of the house also play an important role in prediction of house prices. The housing prices are dependent on the Neighborhood Characteristics. While choosing a house to live in every individual thinks of the surroundings and the locality of the house. In general, houses which are close to shopping malls, Schools, Work locations, and restaurants etc. are preferred over houses located at deserted place. From the available data, we have plotted the different neighborhood characteristics against the sale\_price and drew some conclusions about their relation.

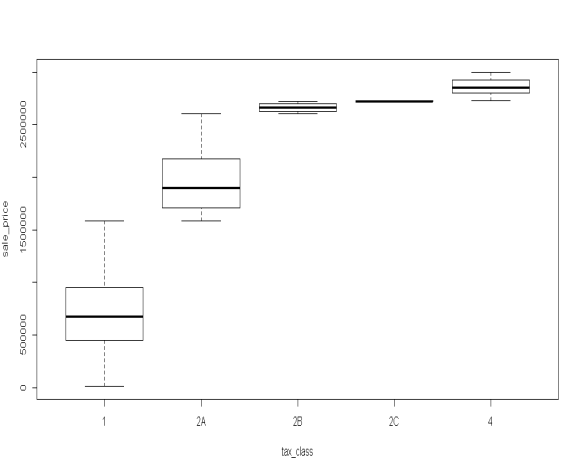
The following is a list of neighborhood characteristics variables in our data.

Neighborhood, tax\_class, building\_class , school dist, council, police pcrt, health cent, easements.

We observed the following results after performing EDA of the variables mentioned above.

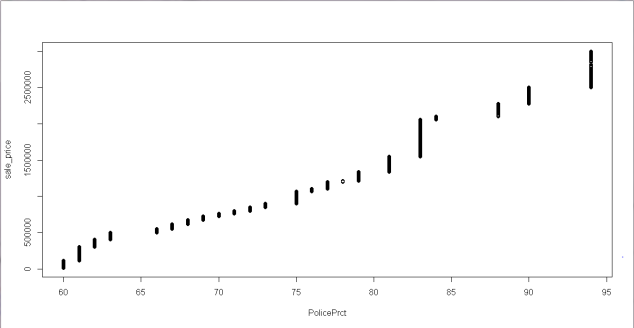


From the graph it can be inferred that the house prices in the Windsor terrace and Red Hook neighborhoods are costlier and the house price with bath beach surroundings are cheaper. The price of houses in these surroundings ranges from a nominal value to very high values.



**Tax class:**

Houses which are categorized in the tax class 4,2C and 2B are priced very high. Whereas houses in tax class 1 record low comparatively prices.

**Police pcrt:**

Sales price is highest in Brooklyn North having Police pcrt in (088, 090, 094), while the sale price is lowest in Brooklyn south area.

# Data Modification

**Feature Engineering**

Based on our observations, we found that there were many outliers which would not be contributing to the model. To remove the unnecessary values, data treatment and feature engineering was done. We made the changes on the following variables

**Removing outliers**

|  |  |
| --- | --- |
| **Variable** | **Filtered and considered values** |
| Sale\_price | $10000-$3000000 |
| LandSqft | 1000-7000 |
| GrossSqft | 100-10000 |
| BldgFront | Less than 200 |
| Residential Units | Below 8 |
| Numbldgs | Less than 4 |
| NumFloors | Less than5 |
| UnitsRes | Less than 9 |

**Dealing with Categorical Variables**

* GarageArea- As we observed that most of the buildings do not have garages, we grouped the variable into 1, i.e. have garages and 0, i.e. don’t have garages
* CommercialUnits – the variable was grouped into 1, i.e. commercial units present else 0, i.e. commercial units not present
* Year\_built – If the year built was between 1850 to 1900, then 1, if it was built between 1900 to 2000 then 2 else after 2000 is 3

**Combining Attributes**

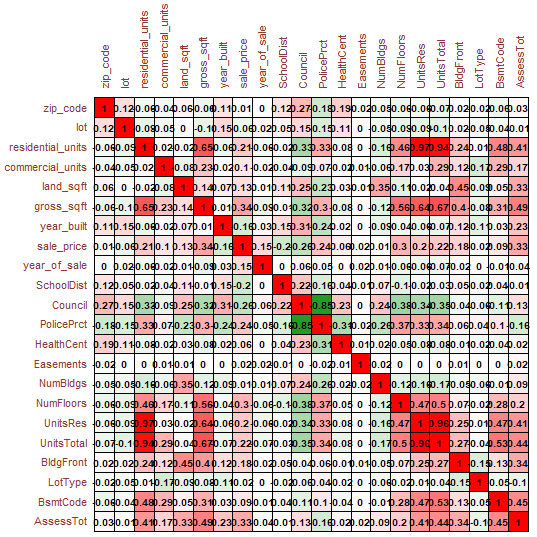
* YearAlter1 and YearAlter2 – Instead of keeping different random variables, we grouped these together to 0, i.e. not altered, 1, i.e. altered once and 2, i.e. altered twice

**Correlation check:**

Next, we proceeded with a correlation plot to find if there was a high relationship amongst the variables. A high correlation means that high values of one are associated with high values of the other, and that low values of one are associated with low values of the other, hence this would impact the variables have on the outcome. After doing the correlation plot, we checked the VIF for variables having high correlation.

We observed that the variable residential unit has a high correlation with UnitsRes and UnitsTotal and gross\_sqft. The variable Units\_Res also has a high correlation with gross\_sqft and unitsTotal. On checking the VIF of these variables, we found it high. We took VIF cutoff as 4 and proceeded with removing the variables residential units and Units\_res

**Initial correlation plot**



|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Normalizing and Standardizing the data**Log Transformations** The main reason why we use log transformation is to reduce skewness in our data. However, there are other reasons  why we log transform our data:   * Easier to interpret patterns of our data. * For possible statistical analysis that require the data to be normalized.   Skewedness:   * A skewness of zero or near zero indicates a symmetric distribution. * A negative value for the skewness indicate a left skewness (tail to the left) * A positive value for the skewness indicate a right skewness (tail to the right)   Kurtosis:   * Kurtosis is a measure of how extreme observations are in a dataset. * The greater the kurtosis coefficient, the more peaked the distribution around the mean is. * Greater coefficient also means fatter tails, which means there is an increase in tail risk (extreme results)   We have numerical variables such as land\_sqft, gross\_sqft, BldgFront, AssessTot. When we checked the skewness of these numerical variables with cutoff 0.8, we saw that the skewness of these variables is high. Regression has an assumption of multivariate normality. It means that regression requires all its variables to be normal. By having skewed data, we violate the assumption of normality.The kurtosis values are also high for mentioned variables. We checked the kurtosis with cutoff absolute 3.    Before Transformation:      We tried Cox-Box transformation to transform the skewed numeric variables for optimized value of lambda. We  transformed the variables using log transformation as well. The variables containing 0 are transformed using  “log (1 + value)” and the variables, not containing 0 in entire range are transformed using “log(value)”.  After Transformation:              Building front still has high kurtosis. We checked for the high values of building front and probability of it being outliers.  **Standardizing the data**  **Scale**: The scale transform calculates the standard deviation for an attribute and divides each value by SD.  **Center**: The center transform calculates the mean for an attribute and subtracts it from each value.  **Standardize**: Scaling and centering help to standardize the data; mean = 0 and SD =1.  We used ‘caret’ package in R which has ‘preProcess’ function to standardize the numerical attributes. We standardized  below attributes:   * + 1. Gross Sqft     2. Building Front     3. Assess Total     4. Land Sqft   ‘predict’ function in caret helps to apply the transformation to original dataset. We merged the dependent variable (Sale price) and transformed numerical data after this step. One-Hot Encoding: Categorical Data One hot encoding creates new (binary) columns, indicating the presence of each possible value from the original  data. It transforms the categorical variable(s) to a format that works better with classification and regression algorithms.  Regression models treat all independent variables as numeric in its purest form. In our case, we have multiple columns  which are categorical in nature. E.g. Neighborhood is a categorical in nature and it should not be translated to  ‘categorical values’, assigning a numerical value to each neighborhood.  The values in the original data are *Red*, *Yellow* and *Green*. We create a separate column for each possible value.  Wherever the original value was *Red*, we put a 1 in the *Red* column.  We used ‘dummyVars’ function in caret package to transform categorical columns to binary columns. ‘predict’ function in caret is used to merge the new binary columns back to categorical dataset.  **Transforming output variable**  Without transformation: After transformation:    The dependent variable ‘Sale Price’ was skewed to the right and is not normalized. The variable does not show normal distribution after using log transformation, but after using square root transformation the skewness is reduced and the density plot shows a normal distribution as shown above. Modeling Model Preparation:  We used two approaches for model data preparation.   * 1. Train-test data split   2. K fold CV   **Model Tuning (Feedback Model Improvement):**  Improved Result  Linear Regression Model  Input data  Remove leverage  Cook’s Distance  We used **cook’s distance** to remove the leverages and feedback the data to improve the model performance and create  the new dataset which we used for building other models. (a)Linear Regression: We created a multiple linear regression model as a base model in R using all the retained numeric and categorical variables. As stated in the above part, the retained variables include scaled numeric variables and modified categorical variables through one hot encoding. We created a custom function in R to measure the adjusted RSquare of the validation (test) data.  We used the data after removing the high leverages using cook’s distance formula as our new dataset. We split in 60:40 training validation dataset. We used ‘sqrt’ transformation for dependent variable ‘sale\_price’ as it was right skewed. We squared the output of the prediction method while comparing it with sale\_price of the validation dataset.  Validity of linear regression model:      As we can see from adjoining figures   1. The residuals don’t form any visible pattern in ‘residuals vs fitted’ graph validating an assumption of homoscedasticity 2. Normal Q-Q plot validates the assumption that all variables to be multivariate normal.   The linear regression model gave an adjusted RSquare value of 68.87% of validation dataset. Running the same model might change the result if seed is not set. (b) Regularized regression: Lasso and Ridge: We used regularized regression model viz. Lasso (L1 regularization) and Ridge (L2 regularization) to check and improve model. We used ‘caret’ package in R to train the model and check the validation. We used below parameters to train the model:   |  |  | | --- | --- | | Method | Repeated CV | | Number | 10 | | Repeats | 10 | | Metric | RMSE | | Alpha | 1 | | Lambda | 0 to 1 |   **Lasso:**  We created a sequence of lambdas from 0 to 1, increasing by 0.01. An alpha value is held constant at 1 for Lasso. We used ‘sqrt’ function to transform the dependent variable in the model and squared the prediction result calculated using validation dataset.    **Tuning Parameter:**  As we can from the above graph, the lambda = 0.02 gave minimum RMSE and MAE as well as maximum RSquared value. The Lasso model gave adjusted RSquared value of 68.80% on validation dataset.   |  |  | | --- | --- | | Method | Repeated CV | | Number | 10 | | Repeats | 10 | | Metric | RMSE | | Alpha | 1 | | Lambda | 0.02 |   **Ridge**:  We used lambda = 0.02 that we got from Lasso tuning parameter. An alpha value is held constant at 0 for Ridge. We used ‘sqrt’ function to transform the dependent variable in  the model and squared the prediction result calculated using validation dataset.  The ridge model gave an adjusted RSquared value of 68.69% on validation dataset. (c)Random Forest: We created a random forest model in R. We used 60-40 training validation data split. We ran model with ntree = 600 initially to check the performance of the model with a change in number of trees.  **Parameter Selection:**   1. Number of Trees:   The graph is plotted between the number of trees and error associated with it. We can see the error rate saturates beyond ntrees > 200. The error rate slowly goes down if we increase the number of trees.   1. Number of columns in each iteration:   By default, the random forest takes sqrt (independent columns) to create trees in each iteration.    **Variable Importance:**  The image shows the various columns and their contributions in the model. The contributions dictate how much each column contributes towards the node purity in various trees present in the random forest. The columns with high contribution does not necessarily mean the positive correlation with dependent variable. It can lead to positive or negative effect.  The variable importance graph along with the graphs created in EDA help to decide the impact and direction of impact on dependent variable. (d)Boosted Tree (Xgboost): We used xgboost package in R to create the boosted tree. We used 60-40 training validation data split. The boosted tree was created by gradient boosting technique. We used multiple techniques to fine tune the model. The final model was an improved version of the baseline boosted model. After trying for the multiple parameters, the following parameters were finalized as they gave better results compared to others.   |  |  |  | | --- | --- | --- | | Parameter | Value | Description | | booster | gbtree | tree based model | | objective | reg:linear | Linear regression | | colsample\_bytree | 0.2 | subsample ratio of columns | | eta | 0.01 | Learning rate | | min\_child\_weight | 2 | min number of instances needed to be in each node | | max\_depth | 4 | maximum depth of tree | | alpha | 0.3 | L1 regularization | | lambda | 0.8 | L2 regularization | | gamma | 0.01 | minimum loss reduction | | subsample | 0.8 | subsample ratio of training instance | | silent | TRUE | TRUE - silent mode, F - printing message | | eval\_metric | rmse | Evaluation metric |     We used cross validation to examine our model. We checked with large number (10000) of trees and algorithm gave the best number of decision trees in final model. We used early stopping in xgboost as we were not sure how many trees we need. Once we got the best number of iterations required (5856), we trained our model of this number keeping other things constant, lowering the early stop point.  The boosted tree gave ab adjusted RSquare (70.37%).    Model Comparison   |  |  | | --- | --- | | **Model** | **R Sq. Value** | | Linear Regression Full Model | 68.15 | | LASSO Regression | 68.41 | | RIDGE Regression | 68.22 | | Random Forest | 69.73 | | Boosted Tree | **70.37** |   As we observe, thee best Model is Boosted Tree which gives an R- square of 70.37. Learning Curve The graph shows the Learning Curve for our best model. The Train Error increases with increase in the training data size while the Test Error decreases gradually with the increase in training data size.  The crossover point shows that the data beyond 50% of the dataset would not lead to any significant impact on the model.  **Ways to improve the analysis**  Availability of certain other variables like Natural Disaster-Prone Areas, Unemployment Rate, Crime Rate, Ethnicity can improve the analysis and provide more insights.  We did analysis using GDP data vs Sales Price, as shown in the graph Recommendations1)Recommendations for customer who are investors: The assessed value is a clear indicator of the sales price and the trend is positive linear.  So, the investor can take this into account and for every $2000 increase in the total assessment the sales price increases by $7,50,000, which is around 125% increase of current value for every $2000 increase in assesses total price.  Investors should take in to account that if a place is getting developed residentially for example if there is only 1 residential unit and 2 more units are getting built there, it can increase the sales price by 150% of the current value.  Noticeably the investor doesn’t need necessarily need to spend extra on a building with a Garage.  Also, buying within a building which has 1 floor as of now but is planning to build 3 floors should be able to get 250% of the current sales price. 2)Recommendations for Builders: Since the Sales price of building increases sharply if the building front is increased from 12feet to 40feet and remains constant afterwards. The optimum building front a builder can look for is 40 feet and can fetch a sales price of $3,00,000.   The sale price increases with number of floors within a range of 1 to 4. So, builder can optimize the price at number of floors equal to 4.  A builder doesn’t necessarily need to utilize resources on the availability of garage to increase the output since there is no relationship between the two.  The plot area and sale price follow an almost positive linear relationship between the range of 1000-7000 sq. ft area. So optimal floor area builder can plan for would be 7000 sq. ft.  Also, since the price increases with residential units, builder can look for an area with 6-7 residential units to maximize the sales price.  Appendix:  <https://www.kaggle.com/tianhwu/brooklynhomes2003to2017>    -------------------------------------------------------------------**END**------------------------------------------------------------------------------- |
| |  | | --- | |  | |