NEw YoRk CiTy  
Apartment Rent Price Analysis



Hasibul Islam  
Due: Dec 19th, 2019

# Introduction

The goal is to find NYC apartment prices for all the user (customer) to find a good price of the apartment for rent. On the project, I use Machine Learning algorithm of example is a technique associated with linear regression, logistic regression, svm and decision tree to help to find an easy way to calculate the prize. Using the machine learning algorithms, I will show owners of apartments different prices based on date that include bathrooms, bedrooms and living space.

This will be displayed by graphs and data values that represent real estate prices. This will help customers (The people looking to rent apartment units) to see the various locations. I will categorize the different models of machine learning techniques and compare each result. Using the python on the Jupyter-Notebook I will be able to determine proper values of apartment units. Also, most importantly, the code will display the graphs that represent price models. This will allow the graphs to visualize better ways to rent an apartment unit. Motivation By using Machine Learning algorithms in this project, the consumers will easily be able to find apartments that suits their price. This algorithm will require few different user inputs which is needed to meet their desired requirements for the apartments. The user will be able to input their recommended number of bedrooms, bathrooms and size. In addition, they will also input the price range they would like their apartment to be. From these inputs, the algorithm will use Machine Learning to determine the apartments that are suited to that consumer.

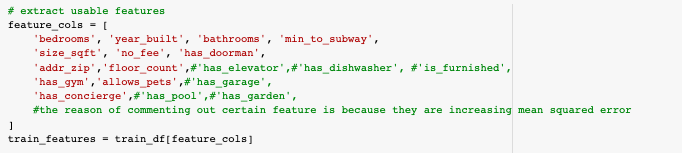
This algorithm along with Machine Learning will also let users see the price ranges in different boroughs. Data Collection I need some data before I can start this project. These data will help with an algorithm to predict the apartments that are suited for the consumer. In this project, i will use the real-estate data.

# Data Collection for Machine Learning

To find the data, I use the nyc open data, in this project i’m using the real-estate data to find to better predict in my data to find an easy way to find good prize. It will help to customer what they are looking for to rent an apartment .

* #### Bedroom
* #### Bathroom
* #### Size in sqft
* $$$$ price for Rent
* New York City-5 Borough

**Adding more feature on this rent price to find accuracy of the prediction.** By observing the data I gathered, the algorithm might favor other boroughs since they have more listings, therefore I limited each to have 100 entries each. The Data shows the raw data gathered.



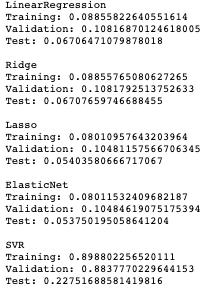
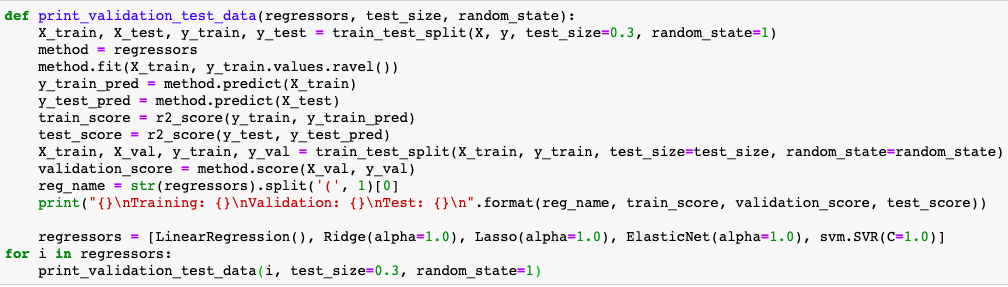
**#DATASET USING IN MACHINE LEARNING**

It will show the better accuracy using algorithm to find a better price for customer what they are looking for rent apartment prize.

# **#First Testing to see how can get prediction the price with different Algorithms**

# Linear Regression

* **1st Test to see the prediction of the price for each Algorithm to predict the price**
* 1st performed got very poorly data and not really correlated
* SVR performed: (22%) than others (6%)
* accuracy( low) and
* Cannot use Linear Regression (22% highest accuracy).



**Final Result of the Test:**

* Linear Regression train data against after test1: 3313817
* Using the Mean squared for test for rent and getting predicted rent price

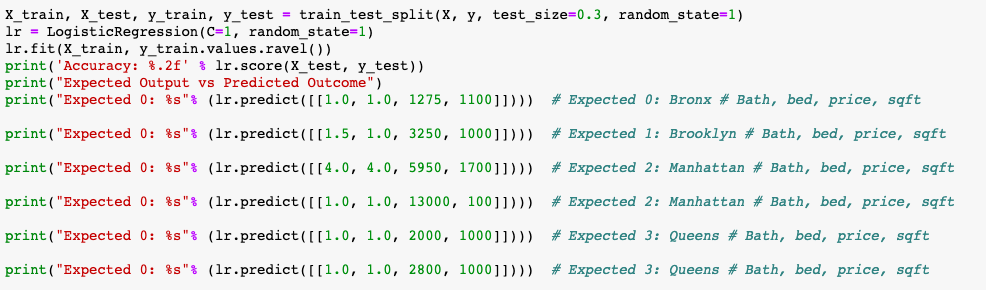


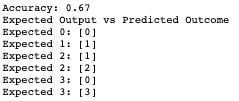
Linear regression is used to predict the value of a variable using another variable. The variable we want to predict is called the dependent variable which relies on the independent variable. Once we know the relationship between these two variables, we can predict the values of the dependent variable. For example, we can say that the test scores of someone is dependent on their study time. People that put more time into studying tends to score higher on tests. Here the independent variable is the studying which the person can control. While the dependent variable is the test score. It looks for relationship but not pre-decided relationship. A relationship is pre-decided relationship when one can be expressed by the other.For example, using temperature in degree Celsius it is possible to predict the Fahrenheit. Since the relationship is not very close in deciding/figuring out the relationship between the two, the error of the distance between the points to the moving backward line. For my project linear regression was the easiest to implement.

# **#First Testing to see how can get prediction the price with different Algorithms**

# Logistic Regression

* 1st Test to see the prediction of the price for each Algorithm to predict the price
* To have better results for 5-borough in DataSet
* Without scaling using algorithm gives correct prediction 67% of the time.
* Better use StandardScaler()



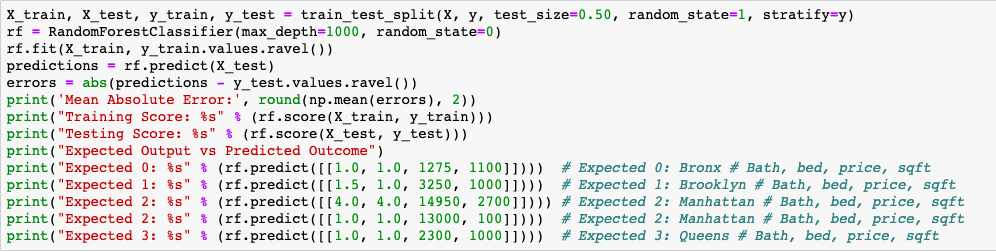


Logistic Regression is another Machine Learning method which uses a classification algorithm to predict the probability of a variable belonging to a category or not. The logistic regression model predicts P(Y=1) as a function of X. Logistic Regression is the most popular way to fit models for categorical data, especially for binary response data in Data Modeling. It is the most important (and probably most used) member of a class of models called generalized linear models. Unlike linear regression, logistic regression can directly predict probabilities; furthermore, those probabilities are well-calibrated when compared to the probabilities predicted by some other classifiers, such as Naive Bayes. Logistic regression preserves the probabilities of the training data. The coefficients of the model also provide hints of the importance of each input variable. Logistic Regression is used when the dependent variable (target) is categorical. This is one method I used for the rent analysis project.

# **#First Testing to see how can get prediction the price with different Algorithms**

# Decision tree

* 1st Test to see the prediction of the price for each Algorithm to predict the price
* Score of 74%
* The algorithm does not learn the data completely (25% error)
* performed 73%



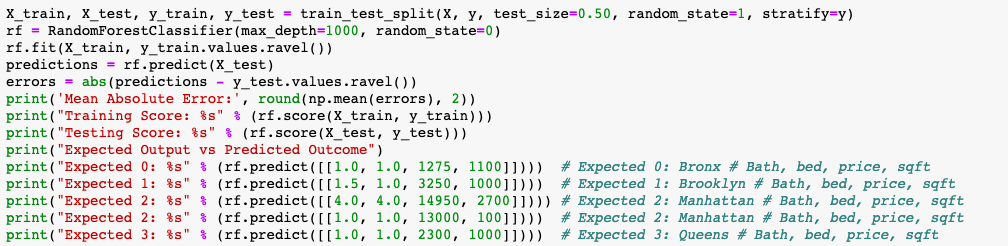


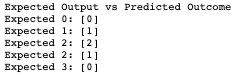
Decision tree models uses tree-like models which contains decisions in order to make a prediction. The tree models that uses the data to classify a target variable is known as Classification Trees. The nodes and leaves of these trees represent the label of the class and the branches are all connections which leads to that class. Regression Trees are also represented in the same method except they are used to predict a continuous value variable. At the end of each node, the tree needs to make the optimal choice and the algorithm which is capable of doing that is known as Hunt's algorithm. This algorithm is both greedy and recursive. It is greedy because it makes the optimal choice at each step. It is recursive because it divides the larger question into smaller questions. The decision to split at each node is made according to the metric called purity. A node is 100% impure when a node is split evenly 50/50 and 100% pure when all of its data belongs to a single class. In order to optimize our model we need to reach maximum purity and avoid impurity. This helped me to understand decision trees and how I should implement in my project.

# **#First Testing to see how can get prediction the price with different Algorithms**

# Random Forests

* 1st Test to see the prediction of the price for each Algorithm to predict the price
* Is very similar to decision trees.
* overfitting by combining series
* data beforehand and it’s not overfitting.
* showed similar score of 74%





* **Final Result of the Test:**

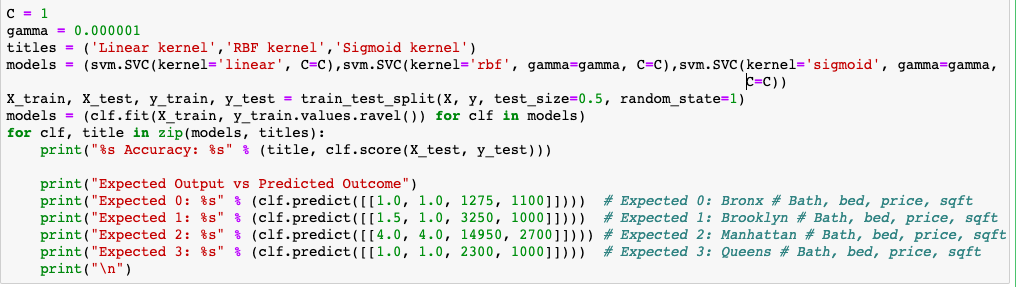


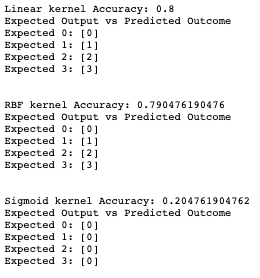
The fourth method is using a supervised classification algorithm known as the Random Forest algorithm. As its name suggests, this method creates a forest by creating a large number of decision trees which are not correlated. The larger the forest is, the more accurate it becomes during prediction. But one thing to note is that creating the forest is not the same as constructing the decision with information gain or gain index approach.The Random Forest algorithm uses decision trees which are tree-like graph to aid in predicting and showing the possible outcomes. This algorithm will create sets of rules which are used for prediction as you input the training datasets, targets and features into the decision trees.If you input a training dataset with targets and features into the decision tree, it will formulate some set of rules. Depending on the data and features you input into the trees, it will create rules to help in predicting the possibilities of something happening. These processes which creates calculates these nodes and creates the rules are using information gain and Gini index calculation. The differentiator between Random Forest and decision tree is in the process of finding the root nodes. In Random Forest method, finding the root node is done randomly. One of the big advantages in using the Random Forest method is that it won’t overfit the classifier if there are enough trees in the forest. I had to make sure that the RFA had enough ‘trees’ to run properly.

# **#First Testing to see how can get prediction the price with different Algorithms**

# SVM

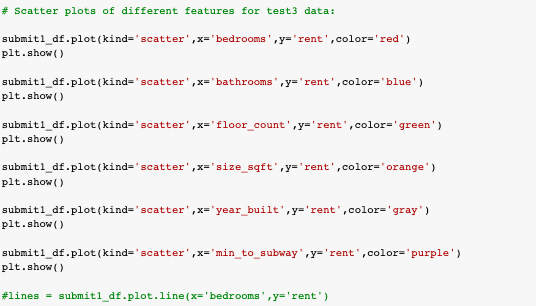
* 1st Test to see the prediction of the price for each Algorithm to predict the price
* linear (80%),
* RBF (79%),
* Sigmoid (20%)
* Get better results

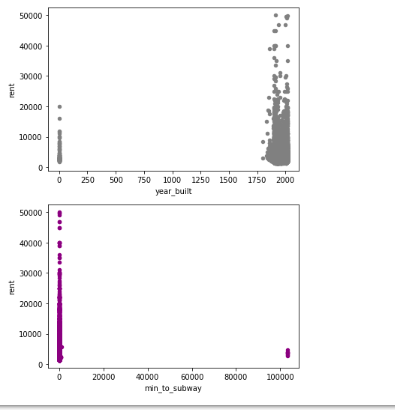
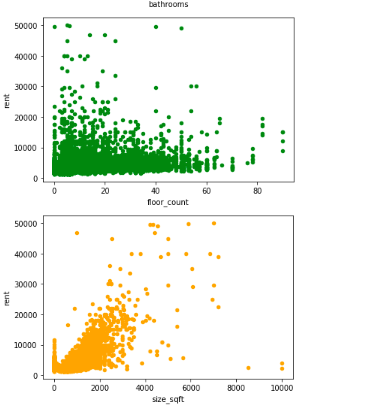
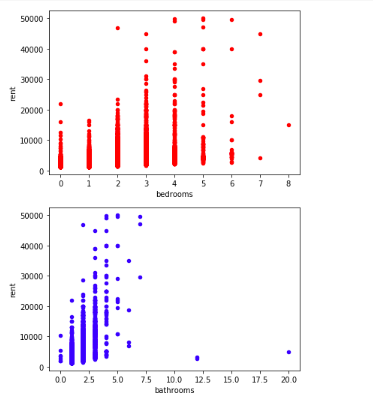


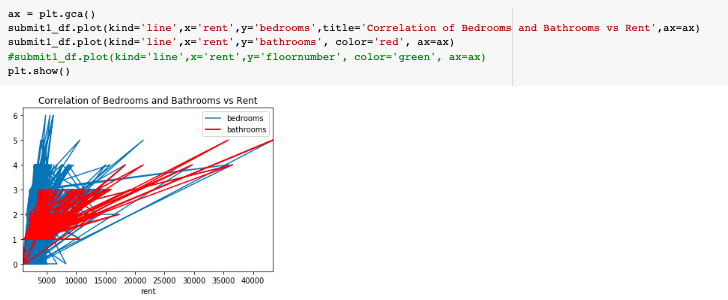
 Another machine learning algorithm is known as Support Vector Machine (SVM). It is a supervised learning model which analyzes data for classification and regression challenges. This is used for classification problems. In this algorithm, I plotted each data item as a point in n-dimensional space with the value of each feature being the value of a particular coordinate. Then, I classificified by finding the hyper-plane that differentiate the two classes. Support Vectors are simply the coordinates of individual observation. It is also a frontier which best puts in two different classes (hyper-plane/ line). The best rule to identify the right hyper-plane: is to select the hyper-plane which segregates the two classes better. This meant I had to make sure the two classes for the project were not contradicting and correct.

# **Result of the Code**

* # Scatter plots of different features for training data:
* Final Result of the Test:
* Feature to find the more rent price to find good prediction price to customer
* bedroom , bathroom, floor count, size sqft, year built, min to subway







# Conclusion

This project has helped me to understand how Machine Learning works and successfully implemented these Machine Learning algorithms in my project. By using the 5 algorithms mentioned above, I was able to obtain an accurate result and accuracy. In this project, I use Linear Algorithm, Random Forests and classification in predicting the price of rent. Other than Machine Learning, I have also learned about Data Science, and the importance of having the right data. Without having the right sets of data, Machine Learning algorithm won’t give us an accurate prediction or even make worse prediction. So, for my project, having the right sets of data about apartments and rent prices, I was able to use Machine Learning algorithm to correctly and accurately predict the rent prices.

short dash