**Prediction of the property price**

**Pranav P**

***JULY 2020***

1. **Introduction:**

Chennai is a metropolitan city in Tamil Nadu, India with a population of about 11 million. According to the Indian census of 2011, Chennai is the sixth-most populous city and fourth-most populous urban agglomeration in India. Recent estimates of the entire Chennai Metropolitan Area economy is of the range from $78.6 to $86 billion, ranking it about fourth to sixth most productive metro area of India. Chennai had been nicknamed the Detroit of India owing to the operation of more than one-third of India's automobile industry in the city. As of 2012, the city is India's second-largest exporter of information technological (IT) products and business process outsourcing (BPO) services. Chennai is a city of dreams for anyone ranging from unskilled labourers to post doctorates and highly educated because of the booming growth of opportunities created by exponential increase of foreign investments in the city. As a proud citizen of this wonderful city I would like to quote a popular saying in our regional language (Tamil).

“Vantharai vazhaveikum singara Chennai”

Meaning, the land of hearts that makes the human existence beautiful. Many people from various cities of Tamil Nadu and other states settle in Chennai and would wish to buy a property that meets all their expectations.

* 1. **Problem statement:**

Chennai is one among the most expensive places to live in India. Buying a property or investing in real estate within the city limits is not a cake walk for relatively new settlers of the area. It is very important to ask around to obtain useful information about the price trends and the safety of the neighbourhood. A good investment in Chennai’s real estate is a great asset, but one can easily be misled by property brokers and end up buying a property that is incredibly overpriced.

* 1. **Interest:**

This machine learning model will be helpful to everyone who is interested in purchasing a property in the most popular places in Chennai. Property brokers might also be interested to monitor the cost trends in real time and also to convince the customers that the predicted price is a fair price.

1. **Data Collection:**

The data leveraged in the machine learning model is obtained from Kaggle. The data can be found [here](https://www.kaggle.com/nishant4k/chennai-house-pricing-). This dataset contains various data corresponding to seven most popular neighbourhoods of Chennai namely Anna Nagar, Chrompet, KK Nagar, T Nagar, Karapakkam, Adyar and Velacherry. This dataset does not include the important amenities around the area. Hence Foursquare API is used to get the neighbouring venues corresponding to the latitudes and longitudes of the area and the price per square feet of the property is predicted based on the data extracted.

For Example: The data collected on the particular area from the Kaggle dataset such as Anna Nagar by adding additional attribute of neighbouring venues the model might be more accurate while predicting prices.

* 1. **Data Description:**

The Chennai house pricing dataset consists of 22 attributes (columns) and 7109 records (rows) in the training set. Whereas the testing set has 21 attributes and 2925 records. Analysing the ‘AREA’ attribute of the training set, it contains seven unique neighbourhood names. The foursquare API is used to obtain the neighbouring venues corresponding to the coordinates of these locations for further analysis.

1. **Methodology:**

House price prediction is a regression problem and since many independent variables has to be taken into account multiple linear regression has to applied in order to obtain reasonably accurate prediction. Cleaning the data is important before performing any kind of exploratory analysis on the data.

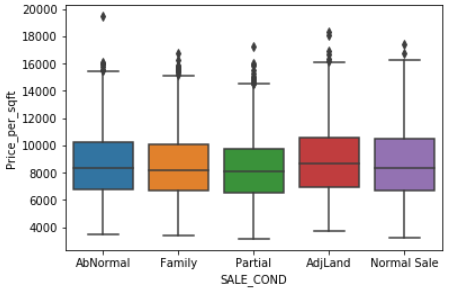
**3.1 Data Cleaning:**

Since the model has to be trained with mostly categorical data, one must encode it before training the model. It is necessary to clean the records to find any misspelt data that might lead to the dimensional increase of the data. The columns are also cleaned for any kind of redundant or missing data.

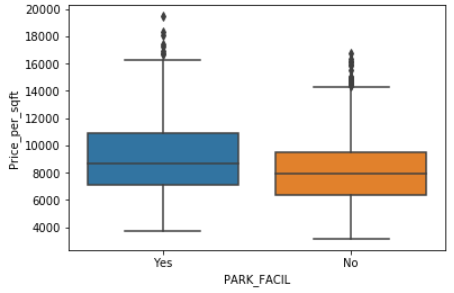
**3.2 Exploratory Data Analysis:**

The certain attributes in the Chennai house pricing data set can be simplified in order to decrease the dimensionality of the model. The attributes such as number of rooms, number of bathrooms and the square feet measurement of the property, registration cost, other expenses and the total cost of the property can be replaced with a single column price per square feet.

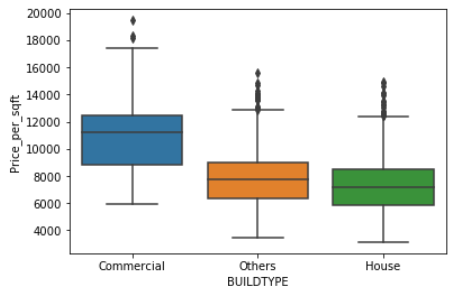
After removing the unnecessary columns the correlation between the column data for each neighbourhood and the dependant variable is measured. Since most of the dependant variables are categorical, box plot can be used. Box plot encompasses important statistical details and gives the user a clear idea of the data at hand. The maximum value, minimum value, median, 1st quartile (median of the lower half of the dataset), third quartile (median of the larger half of the dataset), interquartile range (The distance between two quartiles) and the outliers or whiskers are the statistical details that the box plot depicts. The figure below shows the box plot of ‘Price per sqft’ vs ‘sales condition’, ‘Price per sqft vs type of building’, ‘Price per sqft vs park facility’ and ‘Price per sqft vs street type’.



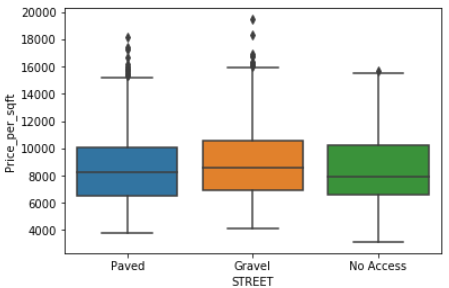
*Fig 1: Box plot sales condition vs price per sqft*



*Fig 2: Box plot park facility vs price per sqft*



*Fig 3: Box plot type of building vs price per sqft*



*Fig 4: Box plot street type vs price per sqft*

The most important categorical variable is the area in which the property is located. The box plot below depicts the relation between ‘Price\_per\_sqft’ and the ‘area’.

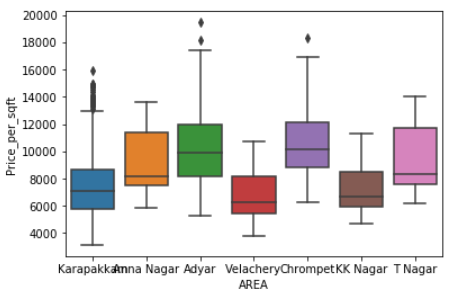


Fig 5: Box plot Area vs price per plot

From the box plot we can infer that only categorical variables Area, park facility and the type of building could actually be potential predictors while constructing the model. Since other categorical variables has significant overlap of their statistical features, including they will not contribute to the precision of the model but will only increase the dimensionality of the problem.

**3.3 Extracting the coordinates of the neighbourhood:**

By making use of geocoders from the geopy library the coordinates of each unique value of the AREA attribute of the data frame is found. The latitudes and longitudes are then plotted using folium for the purpose of visualisation. The found coordinates along with the unique value of the AREA attribute are stored in a new data frame (df\_cord). The two data frames are left joined to form a new data frame.

**3.3 Locating neighbouring venues:**

In Chennai the vicinity from various schools, hospitals and shopping centres is a major factor to be considered while fixing the price. Foursquare API is used for the purpose of locating the nearby popular venues for the corresponding coordinates. An assumption is made for the purpose of analysis. Each venue closest to the neighbourhood is assigned a value weight. For each hospital in the locality the neighbourhood weight increases by two and for each school in the locality the weight increases by one. The weight assigned to each neighbourhood is merged with the main data frame.

**3.3 Encoding categorical variables:**

Categorical variables such as area, park facility and building type encoded using one hot encoder. Once the encoding is done the data frame is ready for training the model.

**3.4 Multiple Regression model:**

The prepared training data is used to train the model by creating the regression object and fitting the dataset to the regressor object. Once the model is trained it is ready to predict the price per sqft for the input data.

**3,4 Prediction:**

The test data is subjected to cleaning, processing and encoding by following the same steps performed to prepare the training data. The predict method is used to find the prediction of the price per square feet.

1. **Conclusion:**

In the study conducted, some additional insights have also been brought to the highlight. From the box plots of built type vs price per square feet the commercial places are often high priced particularly in T Nagar and in Anna Nagar difference between the prices of commercial places and other type of buildings are very high which indicates that these areas are commercial areas whereas other areas are residential areas. From the box plots of area vs price per square feet properties in areas such as Adyar, Anna Nagar and T Nagar are highly priced indicating that these areas are highly popular. This model takes into account the most important features that comes to mind while pricing a property. Property sellers can also use this model to reason the price fixed by them to the buyers. With further improvement and increased precession this model could potentially impact the way in which the real estate business could function.