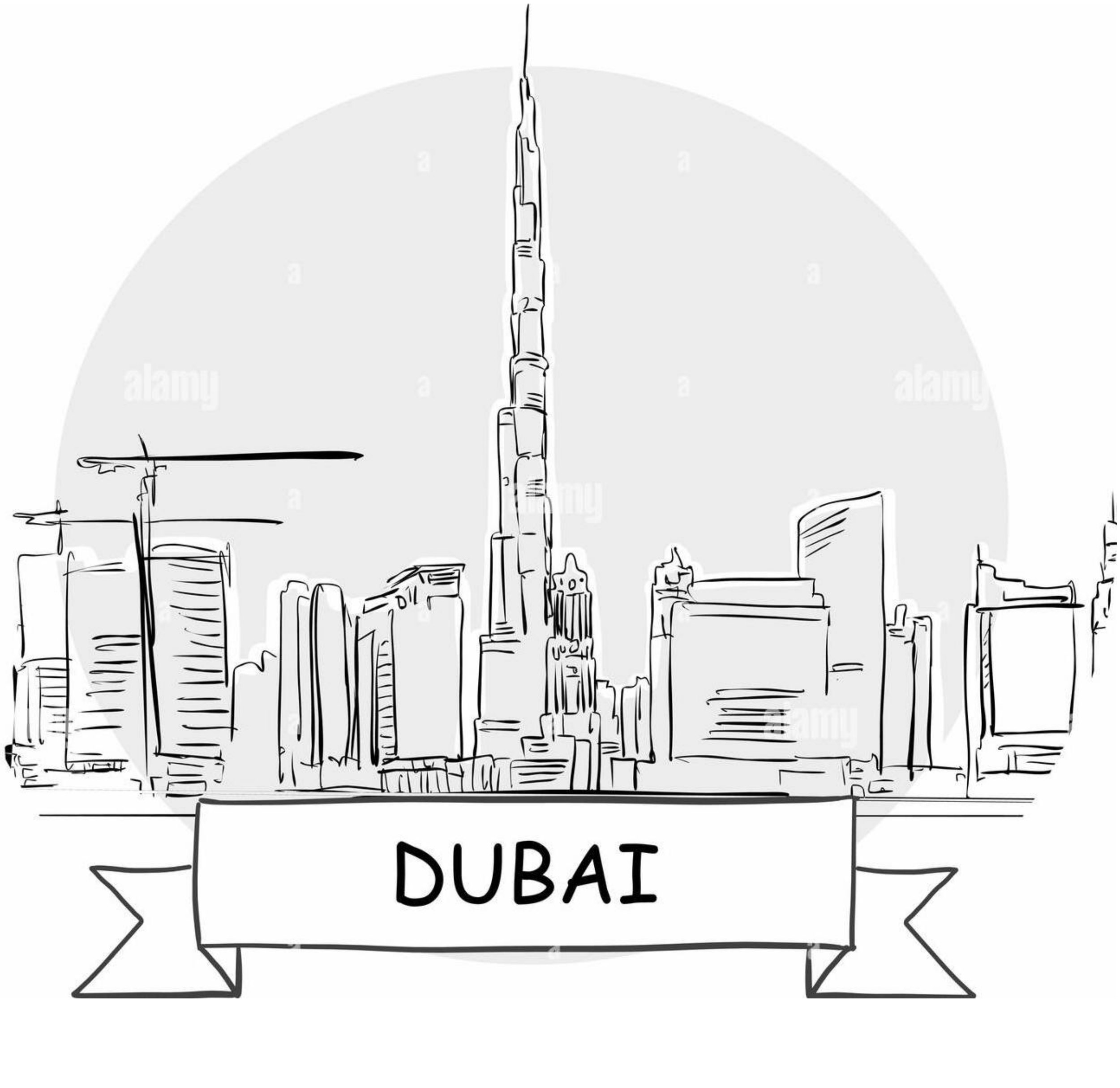
Building a House Price Prediction Model in the City of Dubai

PSTAT 131 - Machine Learning Final Project

Jana Hindiyeh

2022-12-11



Contents

1 Introduction

The aim of this project is to build a machine learning regression model that can predict the price of an apartment in the city of Dubai, United Arab Emirates. I will be using a dataset from kaggle, and testing it in multiple different ways so that it can yield the most accurate price prediction model.

1.1 Why is this model relevant?

Real estate is one of the most important assets that an individual can own, if not the most important. Whether as an investment or as a home or as a place of work, everybody in the world has a connection to, or an interest in real estate. Many different people are interested in many different aspects of this asset, but the most important is of course, its price. That is why I believe that this model is extremely useful and relevant, it could provide me and many others with deep and great insights into the trends behind the real estate market (in Dubai) on a micro level. This model will focus precisely on the sales price of apartments in Dubai based on a host of 37 different predictor variables. This us because this is the type of property that I find most interesting to me and most relevant to younger individuals (since I probably won't be buying a house for a while).

Having lived in Dubai all my life, and worked at a real-tech firm in Dubai, I saw the process of appraising real estate values and thought that it was slightly outdated. A lot of the work was done manually and led to long, tedious tasks that could be simplified using a model like the one I intend to build in this project. I also find the real estate market of Dubai very interesting, since it tends to be a place with a high turnover of residents (especially in apartments), I thought it would be interesting to find out the trends that drive the market.

1.2 Dubai Real Estate Market

There are a few notes I think are helpful to understand about the Dubai Real Estate market. 'New Dubai' or the newer parts of Dubai tend to have a high number of apartments, and is where the majority of younger individuals live. Dubai has a population of 3.5 million people as of April 2022, 3.2 million of those being non-Emaratis ('Emarati' is the term used to define the local population of the United Arab Emirates) or expats. Additionally, around 58.50% of the population is concentrated in the 25-44 age group. What is important to note here is that there is a relatively high turnover of residents that live in Dubai (similar to that of New York City, for example), people that come to Dubai for work and then leave after a few years. Therefore, there is a separation of the demographics that live in apartments and those that live in villas or townhouses. This is why below you will see the map of the properties that are included in this dataset are largely localised in a few areas of the city. These are the areas that surround the main commercial/ work spaces in Dubai, and the rest of the areas are largely populated by houses/ villas/ and townhouses.

1.3 Project Roadmap

Now that a beter understanding of the project and the Dubai real estate market has been established, I want to run through the process in which I will choose the most effective model to predict prices in Dubai.

- Load the necessary packages and the dataset into R.
- Explore the characteristics of, the variables in, and the validity of the data.
- Clean the data and prepare in for use in the model (mutating it accordingly).
- Explore the data, looking at relationships between characters, possible correlations, outliers, etc.
- Model building process. This process will be delved into deeper at the onset of that chapter however, the models that I will be testing are:
  - Ridge Regression

    - Lasso Regression
    - Decision Tree
    - Random Forest
    - ...
- Analysis stage: Analyze models and select best one.
- Test model on Testing set
- Conclusion

1.3.0.1 Loading Packages Firstly, I need to load all the R packages that I need throughout this process.

1.4 Data Introduction

1.4.1 Data source

I found the data on 'kaggle', a website that has a collection of data sets and all kinds of coding projects accross different coding languages. The link to the kaggle website is attached here https://www.kaggle.com/datasets/dataregress/dubai-properties-dataset. The code was sourced in December 2020, therefore, I recognize the limitation of the data regarding the fact that it is not up to date. As a result, its credibility in estimating the property prices as of 2022 may differ slightly.

Firstly, lets read in the data. I downloaded the data as a csv file onto my computer, which is part of the project folder uploaded onto github for your viewing.

```r
house <- read.csv("~/Desktop/PSTAT 131 Final/properties_data.csv")

head(house)
```

Let's look at the dimensions of the data

```r
dim(house)
```

```
## [1] 1905   38
```

The dataset contains **1905** properties and was sourced from a web scrape of the real estate portal. The data contains **38 columns** that consist of characteristics of the properies. Within the data there are a mix of boolean (TRUE/FALSE) variables, integer variables, and character variables, all of which will be used to analyze our data.

1.4.2 Variables of the data set:

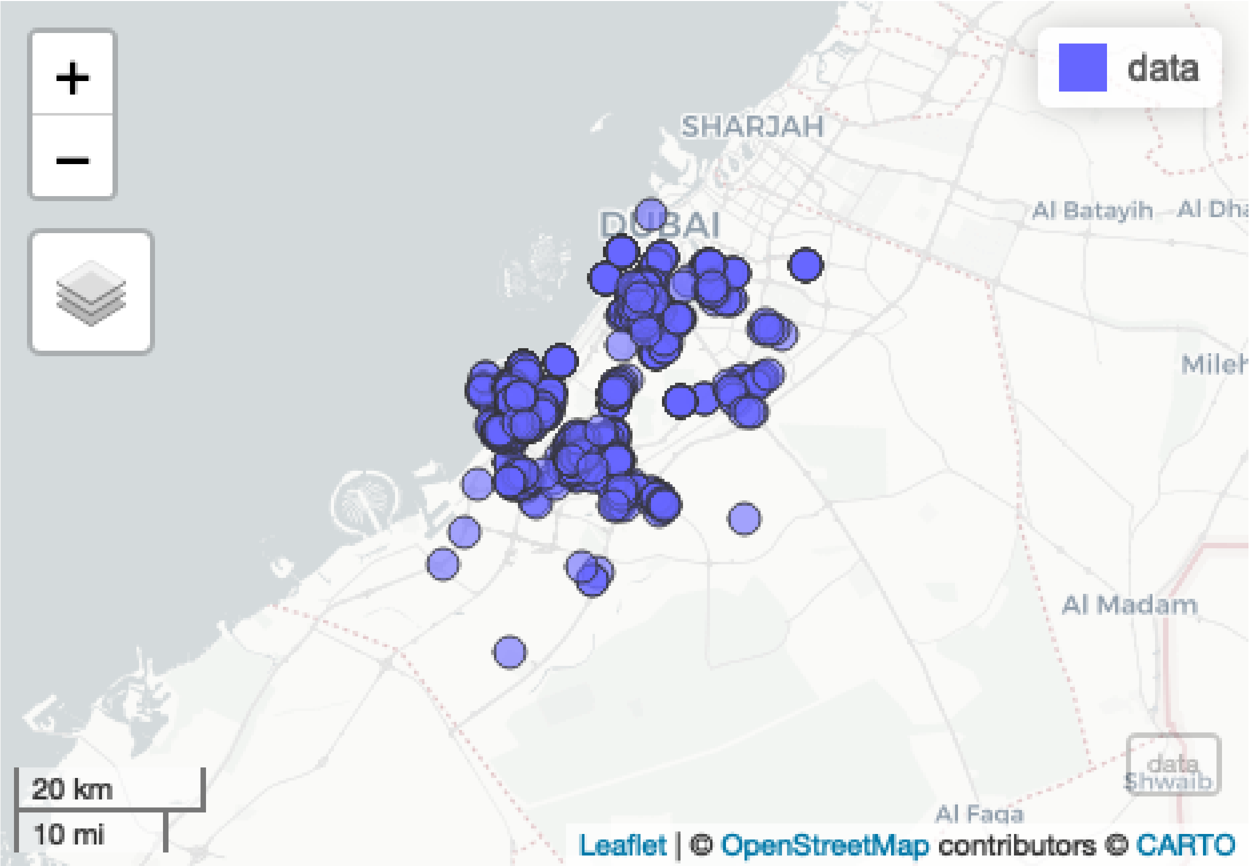
- *id:* Numerical ID for each property.
- *neighborhood:* Neighborhood the apartment is located within.
- *latitude:* Latitude coordinates for the apartment's location
- *longitude*: Longitude coordinates for the apartment's location
- *price:* Property price as of December 2021
- *size_in_sqft:* The size of each apartment in sqft
- *price_per_sq:* Price/ size_in_sqft
- *no_of_bedrooms:* Number of bathrooms in the property
- *no_of_bathrooms*: Number of bedrooms in the property
- *quality:* The relative quality of the apartment (consists of three variables: Low, Medium, High)

- The following are *True or False* observations (e.g., if the property has a maid's room or not)
    - maid_room
    - unfurnished
    - balcony
    - barbecue_area
    - built_in_wardrobe
    - central_ac
    - childrens_play_area
    - childrens_pool
    - …

Note: a full copy of the codebook is available in my zipped files.

Lets also explore the neighborhoods that the data covers bvy reate a map of the data points

```r
map <- mapview(house, xcol = "longitude", ycol = "latitude", crs = 4269, grid = FALSE)
map
```



2 Data Cleaning

Clean the data using clean_names

```
house <- clean_names(house)
house
```

Check for missing values

```
anyNA(house)
```

```
## [1] FALSE
```

Turn all True/False Statements into factor variables:

```
hosue <- house %>%
  mutate(maid_room = factor(maid_room),
         unfurnished = factor(unfurnished),
         balcony = factor(balcony),
         barbecue_area = factor(barbecue_area),
         built_in_wardrobes = factor(built_in_wardrobes),
         central_ac = factor(central_ac),
         childrens_play_area = factor(childrens_play_area),
         childrens_pool = factor(childrens_pool),
         concierge = factor(concierge),
         covered_parking = factor(covered_parking),
         kitchen_appliances = factor(kitchen_appliances),
         lobby_in_building = factor(lobby_in_building),
         maid_service = factor(maid_service),
         networked = factor(networked),
         pets_allowed = factor(pets_allowed),
         private_garden = factor(private_garden),
         private_gym = factor(private_gym),
         private_jacuzzi = factor(private_jacuzzi),
         private_pool = factor(private_pool),
         security = factor(security),
         shared_gym = factor(shared_gym),
         shared_pool = factor(shared_pool),
         shared_spa = factor(shared_spa),
         study = factor(study),
         vastu_compliant = factor(vastu_compliant),
         view_of_landmark = factor(view_of_landmark),
         view_of_water = factor(view_of_water),
         walk_in_closet = factor(walk_in_closet),
         quality = factor(quality))
```