Imperial College London

IMPERIAL COLLEGE LONDON

DEPARTMENT OF COMPUTING

Analysing property sales data using Data Science

Author:
Wenxiang Luo

Supervisor:
Chiraag Lala

Submitted in partial fulfillment of the requirements for the MSc degree in MSc Computing of Imperial College London

June 2022

Contents

Chapter 1

Introduction

Nowadays, there are a significant amount of data generated every second. The daily lives of humans are producing it, and some other fields are generating numerous data, such as research, health care, commercial activities, and environmental information from all kinds of sensors. With these considerable amounts of data, acquiring the relationship between some aspects of records or obtaining the patterns behind these data would benefit the whole world. For instance, new causes of diseases might be discovered, and technology development could be accelerated.

However, this extensive data can be one of the main obstacles for analysis as it is approximately impossible for humans to obtain insights into the data manually. Under this circumstance, artificial intelligence (AI), a technique that empowers the computer to imitate human intelligence and manner, could be one of the methods to mitigate this problem. It can extract patterns from large datasets and use them to make predictions based on future data and even distinguish which parts of data cause the results.

In this project, some AI techniques will be used on property and demographic data to gain insights and understand the factors that impact a homeowner's propensity to sell. The factors could be the distance to schools, hospitals, or supermarkets, the accessibility to public transport, and the types of the properties (flats or houses). It could be highly advantageous to estate agents who would discover homeowners with more potential to become clients and provide them with business.

Chapter 2

Literal Review

Python is one of the most popular programming languages in the world as it is easy to develop and there are extensive packages for various functionalities. In this project, Python and its several packages would be used for loading data, preprocessing data, building and evaluating machine learning models.

2.1 Machine Learning

Machine Learning (ML), a subset of AI, is a technique that the computer can learn and improve from data without explicit programming. The reason for using ML is that its performance is sometimes better than the traditional approach. For example, ML techniques would simplify the solution to a problem that comprises a long list of rules (spam mail detection). ML can be divided into three categories supervised learning, unsupervised learning, and reinforcement learning. In this project, a supervised learning model will be implemented for analyzing data.

2.2 NumPy

Numerical Python (***NumPy***) is a scientific computing package utilizing an optimized C/C++ API to use less computation time than pure Python calculations (McKinney, 2012). It provides some data structures, one of the most important structures is ***ndarray***. Unlike Python lists, NumPy arrays are homogenous, meaning all the elements in them are of the same type, and their sizes are fixed (Fandango, 2017). Moreover, NumPy provides various built-in functions for different purposes, such as statistics, linear algebra, transforms, and element-wise operation.

In this project, ***Numpy*** will be used as part of data preprocessing. For example, converting categorical data into integers as ML models can only use numerical as inputs or standardizing features so that they are in the range of 0 and 1.

2.3 Pandas

The ***pandas*** is a fast and compelling package capable of handling data of various types (numerical values, strings, and time) and from multiple sources (CSV, Excel, and MqSQL database). One of the ***pandas*** data structures is ***DataFrame*** which is appropriate to handle tabular data whose columns are of different types. Also, it could manage various operations, such as manipulating missing values, creating pivot tables (summary of the table), and grouping data from different columns (Fandango, 2017).

In this project, ***pandas*** would be used for data loading and preprocessing. First, data is loaded, and it is represented as ***DataFrame***. Then the attributes of the data, such as distribution, should be inspected. Finally, the data is preprocessed, and one of the steps is handling missing values. For example, they can be dropped or filled with the mean values.

2.4 Scikit-learn

2.5 PyTorch

PyTorch was developed by Facebook, and it is one of the popular libraries for building ML models. It is also the basis of numerous packages that the developers could take advantage of (Godoy, 2021). Moreover, this library could perform auto differentiation by using graphic processing unit (GPU) acceleration, which results in less training and evaluation time.

In this project, some of the components provided by ***PyTorch*** will be utilized.

- ***nn.Sequential*** is the container that contains all the layers of the ML model
- An appropriate loss function would be selected to compute the loss and perform backpropagation.
- One of the optimizers would be used to update the model's parameters.

2.6 Matplotlib

Matplotlib is one of the Python plotting packages that can plot different types of figures, such as histograms, pie plots, and line plots. In this project, this library will be primarily used for data visualization. For instance, the distribution of the raw data should be visualized to determine the procedures of preprocessing and plotting the training loss and validation loss during model evaluation.

Chapter 3

Project Plan

3.1 Data Preparation

1. Obtain the information related to postcodes, such as restaurants and parks. Integrate these data and the resulting ***dataset (A)***, whose primary key is postcode.

2. Download the data, such as accessibility to public transport, for Greater London and then generate the ***dataset (B)***.

3. Download the ***dataset (C)*** for the property whose primary key is the location.

3.2 Data Preprocessing

3.2.1 Handling Missing Values

Some data from the dataset might be missing, and these values should be treated appropriately before passing to the model. In this project, two methods will be attempted to preprocess the missing values.

1. Drop missing values: This method deletes the messing values for data analysis.

2. Fill the missing values: In this approach, the missing values are filled with mean, median, or default values, for example, zero or some constants.

3.2.2 Standardization

3.3 Build Model with PyTorch

3.3.1 Model architecture

3.3.2 Hyperparameters Tuning

3.4 Evaluation

Bibliography

Fandango, A. (2017). *Python Data Analysis*. Packt Publishing Ltd. pages 2, 3

Godoy, D. V. (2021). *Deep Learning with PyTorch Step-by-Step: A Beginner's Guide*. pages 3

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems*. " O'Reilly Media, Inc.". pages

McKinney, W. (2012). *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.". pages 2