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) 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, machine learning (ML) techniques are used on property and demographic data to gain insights and understand the factors that impact a homeowner's propensity to sell. It could be highly advantageous to estate agents who would discover more 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 its extensive packages for various functionalities. In this project, Python would be used for loading data, preprocessing data, building and evaluating ML models.

2.1 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 the size is fixed (Fandango, 2017). Moreover, NumPy provides various built-in functions for different purposes, for example, statistics, linear algebra, transforms, and element-wise operation.

In this project, Numpy is used for 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.

2.2 Pandas

The ***pandas*** is a fast and productive package which is designed to handle tabular data whose contents could be numerical values, strings and time (Fandango, 2017).

2.3 Scikit-learn

2.4 PyTorch

2.5 Matplotlib

Chapter 3

Project Plan

3.1 Data Preprocessing

3.2 Build Model with PyTorch

3.2.1 Model architecture

3.2.2 Hyperparameters Tuning

3.3 Evaluation

Bibliography

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

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