Boston Housing Price Prediction Proposal

March 15, 2023

Group Members: Tony Liang (004), Wanxin luo (003), Xuan Chen (004)

Student Numbers: 39356993, 33432808, 15734643

ECON 323 Quantitative Economic Modelling with Data Science Applications UBC 2023

1 Boston Housing Price Prediction Proposal

1.1 Introduction

In this project, we aim to **explore the impact of environmental factors on housing prices** using the Boston Housing dataset. This dataset contains information on various attributes such as crime rate, average number of rooms, accessibility to highways, and more, which are hypothesized to influence housing prices. The project will involve several parts, including data cleaning, visualization, and model building. Our objective is to conduct exploratory data analysis (EDA) and then build a hedonic regression model with multiple inputs. We will utilize various Python techniques learned in this course to explore the real world data and solve economic questions.

By analyzing the data, we aim to answer economic questions related to the housing market and explore the real-world application of Python techniques. It is important to note that the dataset has its limitations as it was collected almost 50 years ago, but it still provides an excellent opportunity for us to apply our Python skills and gain insights of housing market.

1.1.1 Dataset Description

The Boston Housing Dataset is a derived from information collected by the U.S. Census Service concerning housing in the area of Boston, MA. The following describes the dataset columns:

- `CRIM` - per capita crime rate by town
- `ZN` - proportion of residential land zoned for lots over 25,000 sq.ft.
- `INDUS` - proportion of non-retail business acres per town.
- `CHAS` - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
- `NOX` - nitric oxides concentration (parts per 10 million)
- `RM` - average number of rooms per dwelling
- `AGE` - proportion of owner-occupied units built prior to 1940
- `DIS` - weighted distances to five Boston employment centres
- `RAD` - index of accessibility to radial highways
- `TAX` - full-value property-tax rate per $10,000$
- `PTRATIO` - pupil-teacher ratio by town
- $B - 1000(Bk - 0.63)^2$ where B_k is the proportion of blacks by town
- `LSTAT` - lower status of the population

- `MEDV` - Median value of owner-occupied homes in 1000's

The dataset is derived from https://www.kaggle.com/datasets/fedesoriano/the-boston-houseprice-data.

1.2 Methods

This report strives to be trustworthy using the following steps:

1. Data cleaning
2. Thorough EDA
3. Building multiple linear regression model

Note: this could be subjected to changes later after feedback from the ECON323 Instrutor's team

1.2.1 Data Cleaning

For the data cleaning step, we will check and handle the missing values in the dataset. We will also identify categorical and continuous variables. For instance, the `CHAS` variable is a dummy variable indicating whether the tract bounds the Charles River or not, and is encoded as 0 or 1. Moreover, perform any special treatments toward outliers depending on the method that will be carried in the model fitting phase.

1.2.2 EDA

During the EDA phrase, we will conduct a thorough examination of the Boston Housing dataset. One of the key steps is to generate a correlation matrix, which can help us identify any potential issues related to multicollinearity between the independent variables. In addition, we will use side-by-side box plots to visualize the distributions of the continuous variables and detect any potential outliers or anomalies. Moreover, we will leverage other data visualization techniques, such as scatter plots and histograms, to better understand the relationships between the variables and explore potential trends or patterns in the data. Overall, the goal of EDA is to gain insights into the data and inform our subsequent modelling steps.

1.2.3 Model Fitting

In the model fitting phase, we will split the Boston Housing dataset into training and testing sets. We will then use the training set to select the relevant variables and build our final multiple linear regression model. The selection process can involve various techniques, such as stepwise regression or regularization, depending on the specific requirements of the project. Once we have the final model, we will use the testing set to evaluate its performance in terms of mean squared error (`MSE`). The goal is to ensure that the model can generalize well to new, unseen data and make accurate predictions.

1.3 Division of Labor

Based on the previous discussions, the team has divided the responsibilities as follows:

- Tony: Coding
- Wanxin: Coding and some textual descriptions
- Xuan: Written section of the report

However, the team may make adjustments to the division of labor as needed during the project to ensure that all tasks are completed efficiently and effectively. Effective communication and collaboration within the team will be critical to ensure that everyone is working together towards the same goal.

1.4 References

Vishal, V. (2017, October 27). Boston Housing Dataset. Kaggle. Retrieved March 14, 2023, from https://www.kaggle.com/datasets/altavish/boston-housing-dataset