Analysis of the Housing Market in Two US Cities from 2020-2022

Real Estate Investment Group

**Data Science Capstone Project   
Predictive Modeling Report**

Date:

01/17/2023

Team Members:

Name: Nikhil Muthuvenkatesh

Name: Seyi Oyesiku

Name: Luqing Qi

Name: Rohan Ukkalam

[The purpose of this report is to describe the predictive modeling on the data that you have acquired, pre-processed, and explored in DSCI 591.]

1. Define the Predictive Modeling Problem

1. Input: What are the input data and define the input data clearly?
   1. The input data into the final model would be a real estate listing, which would include the variables <location, beds, baths>, and year. Ideally one would be able to look at the predicted price for a real estate listing depending on what year.
2. Data Representation: What is the data representation?
3. Output: What are you trying to predict? Define the output clearly.
   1. We are trying to predict the price of real estate in Philadelphia and New York City in the future. The output will be a list of prices for the properties that are input into our model.

2. Predictive Models

1. What are the methods? Give a general introduction of the methods with references
   1. Multiple Linear Regression
      1. This model allows us to use multiple features to find out how much they influence the value that we are trying to predict.
      2. The coefficients for each feature in the equation allow us to show what influence that each of the features have on the price. We are also able to notice what value the dependent variable is at each value of independent variables.
      3. Source: <https://www.scribbr.com/statistics/multiple-linear-regression/>
   2. XGBoost
      1. Extreme Gradient Boosting (XGBoost) is an ensemble machine learning method that uses multiple decision trees that are able to correct the prediction errors that previous decision trees have made.
2. Describe the methods with a pseudo code using the definitions in Section 1.
3. Justify the choice of the method.

3. Evaluations

1. What metrics do you use for evaluation?
   1. Evaluation of the final model would partly be conducted via splitting our current data into training and testing datasets for our models. Training the models on the training set we would use the testing set and check to see if the predicted prices are close to the actual prices. The second round of evaluation would be taking a dataset of real estate listings, with the years listed sometime in the future, and comparing the predicted prices for those real estates with research and predictions done by other parties to see if our models are producing reasonable predictions.
2. What is your ground truth?
3. Discuss the performance and the limitation of the method.

**Appendix**

[Addition materials that are not included in the above sections.]

In the first predictive modeling report due in Week 5, you only need to use one method in the predictive modeling. Through the experience of implementing and testing the predictive modeling method, you may learn how it works and fails. In the final predictive modeling report due in Week 10, you will need to try more methods and compare with the one you use in the first report.

Table of Contributions

The table below identifies contributors to various sections of this document.

|  | **Section** | **Writing** | **Editing** |
| --- | --- | --- | --- |
| **1** | **Predictive Modeling Problem Definition** |  |  |
| **2** | **Predictive Models** |  |  |
| **3** | **Evaluations** |  |  |
| **4** | **Appendix** |  |  |

**Grading**

The grade is given on the basis of quality, clarity, presentation, completeness, and writing of each section in the report. This is the grade of the group. Individual grades will be assigned at the end of the term when peer reviews are collected.