CISC351-Assignment1

Jan 2022

1 Introduction

The goal of this assignment is to analyze 3 years'(2018-2020) house sales data provided by New York City (NYC) goverment and build regression model to predict house price. NYC has five boroughs, i.e., Bronx, Brooklyn, Manhattan, Queens and Staten Island. Sales of houses in each borough has been provided. Your goal is to analyze the trends of house sale price in two boroughs and perform prediction of house price in 2020 (randomly sample 10% for testing).

Your submission will be evaluated based on the quality of your analysis, the performance of your model on the testing dataset, as well as the clearness of your textual answers to questions. You can write it in R or Python.

Reference/Tutorial (your can reuse code from online resources, but do not copy and paste with each other, this is an independent homework):

- https://www1.nyc.gov/assets/finance/jump/hlpbldgcode.html
- https://towardsdatascience.com/predicting-housing-prices-with-r-c9ec0821328d
- https://www.kaggle.com/sahilrider/learn-regression-nyc
- https://rpubs.com/jhofman/nycmaps
- https://rpubs.com/ablythe/520912

Submission: you need to submit your analysis as a runnable R notebook or Python Jupyter Notebook. Specify your answer to each sub-question in the markdown (text cell), e.g., Answer to rq1.1.

2 Part 1: Feature Selection and Multicollinearity Analysis

RQ1.1 Random pick 10% of the house price data for the two selected boroughs in 2020 as the testing dataset. The rest data, including those for year 2018, 2019, can be used for training. Note that you can choose 90% of the house price data for 2020 as training. If you believe that 2018 and 2019's data can also contribute to the prediction performance, you are welcomed to use them.

RQ1.2 Analyze the raw features, and pick features you believe can contribute to house price prediction. You need provide statistics or image supporting your claim (e.g., why a particular feature can contribute to house price prediction). Transfer raw features into more meaningful features if needed.

RQ1.3 Detect if multicollinearity exists in selected and newly created features.

RQ1.4 Write a short summary of features you created/used/transfered and filtered in the above steps.

3 Part 2: Prediction using Regression Models

Using the given data set, create a model to predict house price in the testing data. Build at least two regression models based on your analysis from part 1 and measure the performance of your models on both training and testing dataset using Root Mean Square Error (RMSE) metric. Compare the two models and report your findings (which one performs better).