Summary

Yan Hao

Feb 28, 2021

1 Program Design

This program is designed for predicting the price for each airbnb apartment. The program will process in four steps.

First, drop all observations with missing values, extract valuable variables, regroup prices to 7 groups, count how far was the last review till now (last-review can be seen as an activity sign.), convert room_type and neighborhood_group to dummy variables, and then save the clean dataset to full_clean_dataset.csv.

Second, select a suitable model, among linear regression, lasso, and ridge, to predict apartment price. Note, the data used for prediction contain dummy variables of price range. The R square can be very low, around 10%, if we predict prices without price range column. And R square is around 60% by adding price range column. Since Lasso model performs slightly better than the other two, the program saves the trained Lasso machine to a pickle file, called price_model.pickle.

Third, since price range is not included in the raw data, I need to train a machine to predict price range first. Since price range is categorical, the potential models can be used are random forest (RF), logistic, and support vector machine. The latter two models take much longer time to process the prediction, and they do not return a higher accuracy score. Hence, I did a grid search for random forest by changing the number of trees and the maximum depth. The trend can be find in acc_trend.png where x1 to x4 are the four accuracy score sequences, since I split the data to four groups using KFold. The vertical axis presents the actual value of accuracy score. The horizontal axis presents the indexing of the model. This indexing can help you to find detailed information of RF with particular parameter settings in RF_grid_search_results.csv. From the trend in Figure 1, accuracy score of the second and the third splits are clearly higher than the first and the fourth split. Also, the accuracy scores are seasonal fluctuate. The accuracy score goes down as the increase of maximum depth. Hence, I pick n_estimators=16, max_depth=11. This combination gives me a better performance among others. Then the model is saved to PriceRange_machine.pickle.

Fourth, test the model. Now the program has saved the price_model and the PriceRange_machine. I use the first 7000 observations to test the model. In this step, the program firstly drop price and price range columns to form a manually-made test data. Then it predicts price range by using PriceRange machine, and merge this column with the data. And then the machine predicts price using price model. The R square is about 0.2. This is not a good performance, but it is higher than 0.1, which is the R square if we directly predict price without price range. Note, if we can improve the performance of the prediction for price range, the model can perform much better.

2 Problems of this program

First, this program do not use information from latitude and longitude. I tried to find the relation between price and location information. I have plotted each apartment's latitude and longitude colored by different price range in Figure 2. It seems that apartment with same price range are clustered. But I do not find a method to describe it.

Second, since I use pandas to convert dummy variables, if the exist data does not have all price range, then the data used for price prediction will have less column(s).

3 Figures

Figure 1: Accuracy Rate Trend

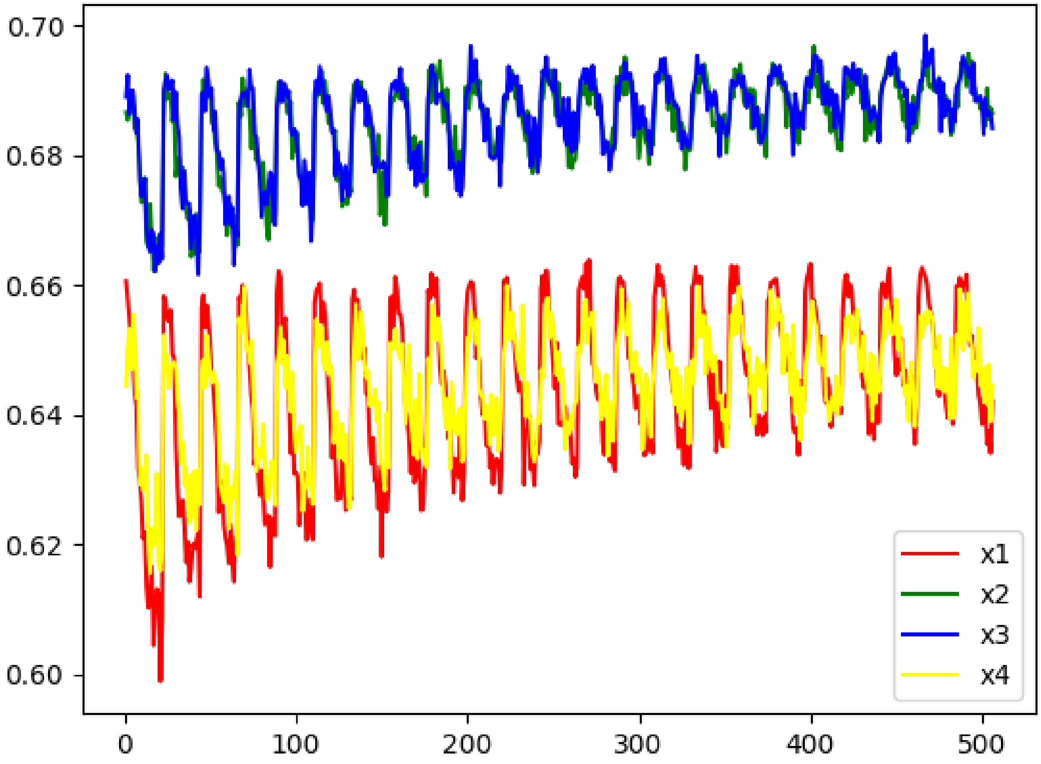
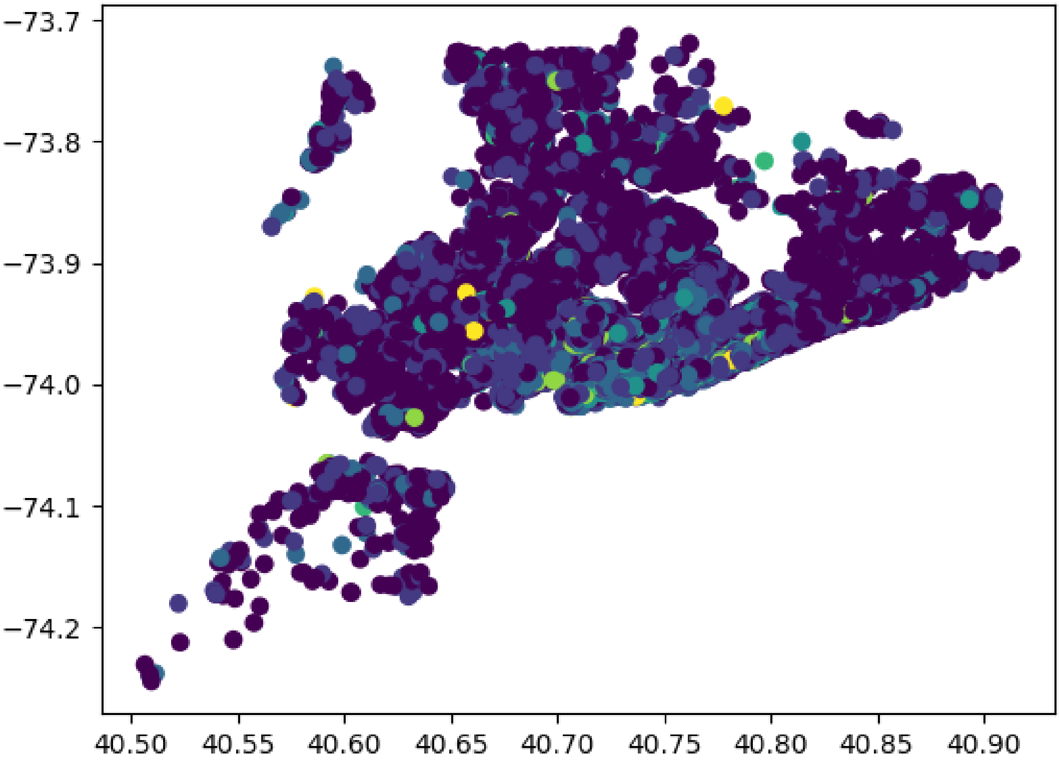


Figure 2: Price and Location


−73.7
−73.8
−73.9
−74.0
−74.1
−74.2
40.50
40.55
40.60
40.65
40.70
40.75
40.80
40.85
40.90