AY2020 Sem 2 BC2407 Computer Based Assessment

HDB Price Prediction with Advanced Predictive Techniques

Introduction

Purchasing a home is the most important and substantial financial comittment for most citizens in developed nations. This assignment will evaluate the performance of advanced predictive techniques to predict HDB prices and elicit the important predictor variables, given 94,373 data points transacted in the last few years (dataset: resale-flat-prices-201701-202103.csv).

A data dictionary is provided in Appendix A. The target variable is resale_price.

One of the advanced technique used is Random Forest. A recent report is included as a PDF file, and can also be read/downloaded here.

Part A: Data Exploration and Preparation (30%)

1. Data Preparation Part 1:

   a. Import the csv dataset as **data1** and ensure that all textual data are treated as categories instead of text string characters. Show your code.

   b. Create a new derived variable **remaining_lease_yrs** (defined as remaining lease in years) from remaining_lease and save as an integer datatype column in data1. Show your code.

   c. Remove lease_commence_date and remaining_lease from data1. Show your code.

   d. Create a new derived variable **block_street** by combining block and street information (with one white space as separator) and save as a categorical datatype column in data1. Remove block and street_name from data1. Show your code.

2. Data Exploration:

   a. Which month year has the (i) lowest transaction volume, (ii) highest transaction volume, and what are their number of sales?

   b. Which town has the (i) lowest transaction volume, (ii) highest transaction volume, and what are their number of sales?

   c. Generate an output that shows the top 5 resale prices and bottom 5 resale prices in terms of flat_type, block_street, town, floor_area_sqm, storey_range, and resale_price.

   d. Conduct additional data exploration. Show (with screenshots of software outputs) and explain the interesting findings discovered.

3. Data Preparation Part 2:

   a. Copy data1 and save as data2. Show your code.

   b. Remove flat_type "1 ROOM" and "MULTI-GENERATION" cases from data2, and ensure these levels are also removed from the categorical level definition[1]. Show the categorical levels of flat_type and list the number of cases by flat_type.

   c. Remove block_street from data2. Show your code.

   d. In data2, create a new variable **storey** by copying storey_range, and then create and use the categorical level “40 to 51” to combine all the relevant storey levels into this bigger category. Show and verify that the categorical levels in storey are created correctly to hold the right cases.

   e. Show the categorical levels in storey and list the number of cases by storey.

   f. Remove storey_range from data2. Show your code.

   g. Remove flat model "2-room", "Premium Maisonette" and "Improved-Maisonette" cases from data2, and ensure these levels are also removed from the categorical level definition. Show the categorical levels of flat_model and list the number of cases by flat_model.

   h. How many cases and columns are in data2 after completing all the data prep steps above?

   i. Suggest a reason for executing such data preparation steps listed above.

[1] Removing all cases with specific categories will not remove these categories from the categorical level definition. Example: Your dataset may be subsetted to contain only males, but the categorical level definition for gender column in the dataset could still have both male and female categorical levels defined. In R, the categorical level definitions for a factor variable can be checked via the levels() function. Ref: https://techvidvan.com/tutorials/r-factors/

In Python, the Pandas library provide several ways to check, create and remove categorical levels. See https://pandas.pydata.org/pandas-docs/stable/user_guide/categorical.html

Part B: Advanced Predictive Techniques and Insight (40%)

4. Set seed as 2021 and do 70-30 train-test split on data2. Execute (i) Linear Regression (all predictor variables), (ii) MARS degree 2, and (iii) Random Forest to predict the target variable. Create and show a summary table that lists the trainset RMSE and testset RMSE for the 3 models (to nearest dollar). Which model performed the best?

5. What is the OOB RMSE of the Random Forest? Can this be used as the estimate of Random Forest performance instead of testset RMSE? Explain.

Part C: Critical Thinking (30%)

6. In Soifua (2018) report [2], Gini-based variable importance was used instead of Accuracy-based variable importance via Perturbation. What are the Pros and Cons of using Gini-based variable importance? Is this better than using Accuracy-based variable importance?

7. Soifua (2018) report Appendix B mentioned the idea of weighting each tree in the random forest by their respective OOB error to improve model performance. However, it was “concluded that the improvements were too modest to be worthwhile…” Suggest a reason based on your understanding of random forest.

[2] Soifua, Breckell, "A Comparison of R, SAS, and Python Implementations of Random Forests" (2018). All Graduate Plan B and other Reports. 1268. This report was provided as part of CBA question paper folder as a PDF and the original website link is given in page 1 of the CBA question paper.

Appendix A

month: Year and month of the transaction.

town: Area location of the flat.

flat_type: Type of flat.

block: An identifier for the exact building of the flat.

street_name: An identifier for the street location of the flat.

storey_range: The level of the flat within the building, given as a range.

floor_area_sqm: Floor area of the flat in square metres.

flat_model: Another classification for type of flat.

lease_commence_date: Start Date of lease. All flats have 99 years of lease at start.

remaining_lease: Remaining “life” of the flat. Upon expiry, the flat has no market value and is returned to the Government.

resale_price: The selling price of the flat in the resale market.