MSBA7027 Machine Learning

Homework 2

Due 11:59 pm Jan. 20, 2024

Notes:

- You are required to submit 1) original R Markdown file and 2) a knitted HTML or PDF file via Moodle. Please provide comments for R code wherever you see appropriate. In general, be as concise as possible while giving a fully complete answer. Nice formatting of the assignment will receive extra points.
- Remember that the Class Policy strictly applies to homework. You are encouraged to work in groups and discuss with fellow students. However, each student has to know how to answer the questions on her/his own.
- Please allow some buffer time and do not submit homework at the last moment. You will have points deducted if you submit the above two items late (even by one minute).
- Please note that to be fair to all students, the instructor and the TAs can only answer clarification questions about the assignment.
- For this homework, you do not need to perform data preprocessing.

**Question 1.** Load the dataset from **HW2_house_dataset.csv**. You will implement some tree-based methods to predict housing prices. Basic characteristics of the dataset are given as follows:

- Problem type: supervised learning, regression
- Response variable: selling price of houses (in log10 units)
- Data variable name in R: "price"
- Number of features: 17
- Number of observations: 21,613
- Task: use house attributes to predict sale price of a house

Please perform the following tasks:

(1) Set seed
(2) Perform stratified sampling, use 80% as training and 20% as testing. Do not touch the testing data until the last problem (6).
(3) Perform random forest (RF) on the training data. Find the best tuning parameters and describe how you find them, and after that report the smallest cross-validated RMSE on the training data. Which four predictors are the most important? Obtain PDPs for these four predictors, describe them and provide possible explanations.
(4) Repeat (3) for basic GBM algorithm.
(5) Are the four most important variables different in (3)-(4)?
(6) Among RF and GBM with their own best-tuning parameters, which one has the smallest cross-validated RMSE on the training data? Choose that method, refit the model with all of the training data, use that model to make prediction on the testing data, report the RMSE for the testing data.

Is the obtained RMSE smaller or larger than the cross-validated RMSE?

**Appendix: Description of Features**

- price (numeric): sale price (log10 units)
- bedrooms (numeric): number of bedrooms
- bathrooms (numeric): number of bathrooms
- sqft_living (numeric): size of living space
- sqft_lot (numeric): size of property
- floors (numeric): number of floors
- waterfront (numeric): binary indicator for a waterfront view
- view (numeric): rating of the quality of the view
- condition (factor): condition of the house (poor to very good)
- sqft_above (numeric): size of living space above group
- sqft_basement (numeric):size of living space below group
- yr_built (numeric): year build
- year_renovated (numeric): year renovated and, if not renovated, the year built
- zip_code (factor): zip code
- latitude (numeric): latitude
- longitude (numeric): longitude
- nn_sqft_living (numeric): size of living space of 15 neighbors
- nn_sqft_lot (numeric):size of lot of 15 neighbors

**Question 2 (Optional, 10 bonus points).** In Machine Learning, we are interested in the relationship between a target variable (dependent variable) with respect to independent variable(s). Identify three such relationships of your interest, where there is a potential causal relation between the independent variable(s) and the target variable.

Requirements

- The independent variable(s) can be continuous, categorical or time-series: representing events, shocks, interventions, change in patterns or status.
- The target variable can be continuous or categorical.

Examples

- Healthcare: [y: #kidney transplants between two cities] versus [x: #direct flights between the two cities]
- Online Platform (Food-delivery): [y: #single-use utensils ordered] versus [x: Indicator variable whether to "green-nudge" (e.g. message that a tree will be planted if utensil is not ordered) the customer]
- Social Media: [y: Indicator variable whether the last digit of WeChat red packet is 8] versus [x: Indicator variable whether the red packet is sent during Chinese New Year]

For any one of the three relationships, acquire relevant datasets and

- Perform a linear regression for the target variable against the independent variable(s)
- Without performing a non-linear model, explain whether you think a non-linear model will perform much more accurately than the linear model.

For this question, you do not need to perform data pre-processing. You will be graded on

- whether the relationship is innovative and interesting
- whether the dataset is unique
- whether the dataset is comprehensive and of good quality