SOUTH UNIVERSITY OF SCIENCE AND TECHNOLOGY OF CHINA

Intro to Big Data Science: Midterm Project

Due Date: Nov 27

✏ **Problem**(Predicting housing price)

House sales are now such a hot topic in China that everyone pays too much attention on their changes. Similar cases happen in USA. We now have a set of house sales data in our hand. This dataset contains 21613 items of house sale prices for King County, which includes Seattle. It includes houses sold between May 2014 and May 2015. We want to predict the house price using these data.

The data has 21 attributes (columns), where the column names have their own meanings. For instance, column 1 is the IDs of the records which seem to be useless and unnecessary in predicting housing price. Column 2 is the date when the houses are sold. Column 3 is the price information. Column 4 is number of bedrooms. Column 5 is number of bathrooms. Column 6 is the size of living rooms in unit of sqrt. Column 7 is the actual area of land occupied by the house. Column 8 gives how many floors in the house. Column 9 gives the number of waterfront in the house. “view” provides how many views in the house. “condition” shows the quality of the house condition, whether it is good or bad. “grade” shows which grade the house is in. “sqft_above” is the area above the land (not sure). “sqft_basement” is the size of the basement. “yr_built” tells when the house is built. “yr_renovated” tells when the house is renovated. “zipcode” lets you know roughly where the house locates. “lat” and “long” indicate the exact latitude and longitude of the house’s location. “sqft_living15” and “sqft_lot15” are the corresponding values in 2015.

For the purpose of exam, the data “kc_house_data.csv” shall be split to a training set (70%) and a test set (30%). The training set contains 15129 house sales records while the test set contains 6484. For simplicity, we just choose the first 15129 items in the “kc_house_data.csv” data to be training data and leave the remaining items to be test data.

You goal is to use the training data to fit a regression model between "price" and other attributes. The performance of you fitted model should be tested by the test data. In this project, you have the following tasks to do:

1. Data exploration: including data statistics and data visualization;
2. Data preprocessing: including split the data into training set and test set, detect missing values (if any) and outlier samples (if any), data conversion and normalization (if necessary);
3. Model construction: you could use any model you prefer, even the model we did not cover in class;
4. Feature selection and model selection;
5. Model evaluation;
6. Conclusion.

These should be included in your report. You should also have a complete set of codes. You report (typically in pdf format) and codes should be compressed in a zip file. Please use your student ID and name to rename your zip file. Then the zip file shall be uploaded to the classroom of hackdata.

Your project will be graded based on several factors, including the accuracy (e.g., Mean square error, R^2 score or adjusted R^2), comparison of different methods, whether your methods are innovative, the quality of your report, the analysis on your methods and you results (e.g., computational efficiency, model interpretability, etc.), and the quality of your codes.

Students may submit their work in the form of team players. Team can be made of 2-3 players. Every team member should demonstrate which part he/she is in charge of in the teamwork. And he/she should also defense on his/her contribution. This is important to reveal the team member's role in the his/her final score.