QBUS2820 Assignment2

This version was compiled on October 30, 2020

Introduction

In a traditional manner, sale prices of houses were predicted by comparing sale prices and costs in the real estate market. There was no general standard to estimate the value of houses. Machine learning techniques therefore play an important role to help establishing models for sale prices of house predictions. As mentioned by Calhoun, the availability of a house price prediction model helps fill up an essential information gap and improve the efficiency of the real estate market (Calhoun, 2003).

This project aims to develop predictive models for sale prices of house with machine learning techniques. With the sale price which is a numerical variable being the response of predictive models, six models are developed and validated.

By comparing the root mean squared errors of predictions, the lasso regression model and random forest model are found to have the best predictive performance for the housing data, compared to elastic net, ridge regression, k-nearest neighbour regression and stepwise regressioin with forward selection.

Data processing and exploratory data analysis

There are 36 numeric variables and 43 categorical variables in the housing data

Feature engineering

As shown in Figure 1, there are huge amounts of missing values in several columns: 'Alley', 'Fireplace Qu', 'Pool QC', 'Fence', 'Misc Feature'. By calculating the number of null values, these 5 columns are found to have more than 40% missing values within each column. With this issue, such variables are uninformative to ba a feature of predictive models as too few observations are provided. Removing all rows with missing values can lead to significant loss of data while imputation is not appropriate for such largely incomplete columns. Therefore, 'Alley', 'Pool QC', 'Fence', 'Misc Feature' are abandoned.

Besides, there are 19 columns containing missing value but the percentages of missing values are less than 20%. This can be dealed with by imputation. The missing values are im-

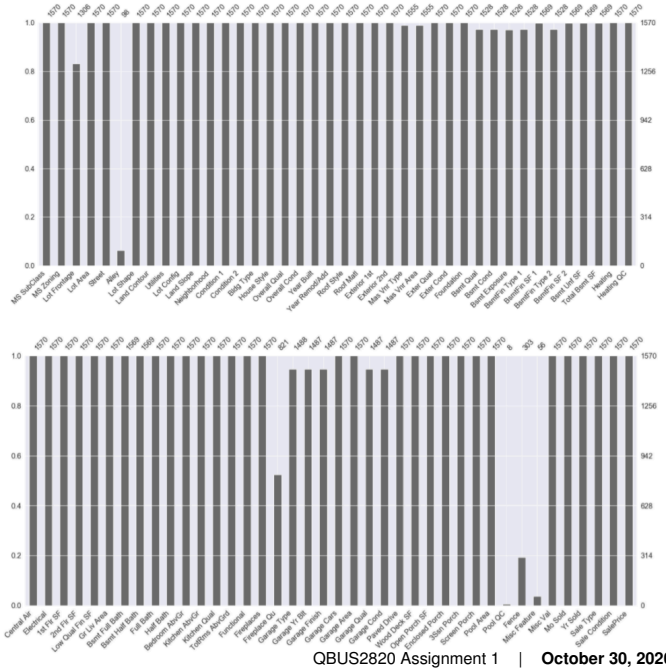

Fig. 1. Visualizing missingness of housing data.

puted by using the most frequent value of each columns.

To deal with outliers of numeric features, standardisation is performed by subtracting the mean, followed by dividing the standard deviation of the corresponding columns.

After feature engineering, there are 74 informative features, with 36 features being numerical and 38 features being categorical. There are 1570 in the training set while 1210 observations remain in the testing set. In order to involve categorical features in regression models, dummy variables are created for each categorical feature.

Methodology

Validation set