STAT 450 Real Estate: Individual Report

yuting wen

4/13/2020

Summary

The main objective of our project is to accurately predict mill rate (property tax) in metro Vancouver for the following 3 property tax classes: 1) Tax class 1 - Residential; 2) Tax class 5 - Light industry; 3) Tax class 6 - Business and other in 2020. Every year, the assessment value of each property is released at the beginning of the year; however, the mill rate is still unknown until Spring. Prediction of mill rate is a focus of interest because it gives an approximate property tax to pay for property owners.

Data cleaning and exploratory data analysis are used to analyze the relationship between mill rate and other factors. Data cleaning is performed to aggregate our data into summary statistics. The exploratory analysis shows that Vancouver, Richmond, Burnaby, and Surrey have much higher total assessment than other municipalities. To reduce the effect of outliers, the average total assessment was calculated by taking total assessment dividing by the number of properties of each municipality and tax class. The exploratory analysis shows a fairly strong correlation between mill rate and average total assessment; it also shows that there are strong relationships between mill rate and tax class and mill rate and municipality.

Ordinary linear model, reduced ordinary linear model, ridge regression, and LASSO are used to predict the mill rate. The ordinary linear model uses all variables from summary statistics after aggregation; the transformed ordinary linear model uses a subset of variables based on the results of exploratory analysis; ridge regression and Lasso are used to optimize the ordinary linear model.

Cross-validation is performed to calculated the mean squared prediction error to measure the performance of different models. We decided to use the transformed ordinary linear model to predict the mill rate in metro Vancouver in 2020 because it has better performance and a simpler form compared to other models.

Limitations and future directions

- We aggregated the dataset into summary statistics to reduce the dimension of our data. Instead, we can fit a model for each municipality and each tax class separately. The models we fitted only have approximately 4 data points for each municipality and each tax class, while we can have way more data points to use if we fit models separately. Although fitting models with more data points usually have better accuracy, our data have a lot of missing values and outliers and bad data pre-processing might cause worse results. Hence there is a trade-off.
- We only compared our models based on the mean squared prediction error. We would've used more comparison methods to measure the performance of our models like AIC, BIC to make our decision more reliable.
- We used our transformed ordinary model to fit the 2020 mill rate, and we found that the mill rates from a particular tax class in two cities are negative, which should be impossible values. We can add more restrictions to the mill rate while fitting models.