Final Report - Individual

Yuetong Liu 43838168

Summary:

The main objective of our project is to accurately predict future mill rate (property tax) in metro Vancouver for the following 3 property tax classes: Tax class 1: Residential, Tax class 5: Light industry and Tax class 6: Business and other. Data cleaning and exploratory data analysis are used in this project to analyze the relationship between mill rate and other factors. Data cleaning is performed to aggregate our data into summary statistics. Exploratory data analysis shows that there are strong relationships between mill rate and tax class and mill rate and municipalities; it also shows there is a fairly strong correlation between mill rate and average assessment per property in different municipalities.

Individual Contribution:

In this project, I mainly focus on data preprocessing and calculating the 2020's mile rate using the chose model. The raw data has around 2 million data entries and 30 features. To reduce the dimension of the data, I select 5 features that could be relevant to the mill rate and all properties are grouped by municipality, tax code and tax year since each of these groups has a unique mill rate.

Main Result:

Transformed OLR, Ridge Regression, and LASSO were able to make good predictions based on mean squared prediction error on the training sets and test sets from cross-validation. Since a simpler model is preferred, we choose the transformed model to make our 2020 prediction.

Observation:

Based on the exploratory data analysis and prediction model, there are three factors (Tax Class, Municipality and average assessment value per property) that show strong associations with the mill rate.

Limitations and Future Directions:

- Since we only have five years of data, it is too short to conduct time series analysis. Therefore, we are unable to analyze the flotation of mill rate by year.
- The municipal budget could be a significant factor that associates with the mill rate because property tax is the major income of the BC government. However, it is not incorporated in our model due to a lack of data sources.
- Our client is interested in implementing a shiny application that can predict mill rates give past property data. However, the property data set contains massive missing data, and they are missing for different reasons (missing at random, missing completely at random). Therefore, multiple imputation methods are applied. This is a judgemental process that could be hard to automate.