STAT 450 Project: Real Estate

Peter Han, Xuechun Lu, Yuetong Liu, Yuting Wen

07/04/2020

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, 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 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. *Ordinary linear model*, *reduced ordinary linear model*, *Ridge Regression* and *LASSO* are used to predict the mill rate. A full assessment of the performance — prediction power and goodness of fit of these models, is shown below.

Introduction

Our goal is to predict mill rates in 2020 in Metro Vancouver.

We also seek to identify which explanatory variables are the most important in determining mill rates. 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. It is also important because it might affect future buyers' purchasing incentives. The property tax rate has a fairly small margin to change. Mill rate is adjusted based on the total assessment in each city so the municipal government can use tax earning (total assessment * mill rate) to match their annual expense to balance the city's budget.

Correlations between the mill rate and each explanatory variable are used to pick the essential variables in our model. Then, a variety of linear models are fitted using our selected variables. The best model is selected based on its prediction power and goodness of fit.

Data Description

Our client provided us with the past 5 years' property assessment data in BC. Since the only interest is in predicting mill rate for metro Vancouver and specific tax classes, a subset of properties that satisfy our interest have been selected:

- Tax Class in (01,05,06)
- Municipality in (Burnaby, Coquitlam, Delta, Langley - City, Langley - Township, Maple Ridge, Maple Ridge Rural, New Westminster, North Vancouver - City, North Vancouver - Dist, Pitt Meadows, Port Coquitlam, Port Moody, Richmond, Surrey, Vancouver, White Rock, West Vancouver, Bowen Island, Anmore, Belcarra, Lions Bay)

Moreover, 5 features have been selected that could be relevant to the mill rate:

- Tax Year
- Municipality
- Tax Class
- Assessment Type
- Assessment Value

There are 1801 missing value in mill rate. 1509 are imputed, and 292 are removed from the data frame. The imputation method is mentioned in **Appendix**.

To reduce the dimension of our data, all properties in the same region, tax class code, and year are aggregated into a group because these properties have the same mill rate, which is our response variable. Here are the summary statistics for these groups:

- Mill rate (rate)
- Total Assessment (assessTotal)
- Total Land Assessment (landTotal)
- Total Improvement Assessment (improvementTotal)
- Total Number of Properties (propertyCount)
- Tax Class Code (TaxClassCode)
- Municipality (AddressAssessorMunicipalityDesc)

Methods

Exploratory Analysis

Before any prediction on the future mill rate of Metro Vancouver's real estate market was made, exploratory data analysis was performed to explore and visualize the main characteristics of our dataset. Correlation analyses between Mill Rate and Total Assessment, Mill Rate and Total Land Assessment, and Mill Rate and Total Improvement Assessment were performed. From our initial analysis, outliers in municipalities were found. Data transformation — calculating the average total assessment, was used to reduce the effect of outliers.

Mill rate was mainly affected by assessment, so scatter plots of mill rate vs. total, land, and improvement assessment were created to see the correlation between each pair of the two factors. Kruskal Wallis analysis was also performed to see the correlation between mill rate and tax class and mill rate and municipality.

Refer to **Reference** for more information on the Kruskal Wallis Test.

Measure of goodness of fit and prediction power

In this study, linear models were built and the performance of each model was evaluated by the goodness of fit and prediction power, that is, how well the model explains the data and how well it can predict future values, respectively. The definitions of the goodness of fit and prediction power are given below.

Prior to examine the goodness of fit and prediction power, the dataset was divided into training sets - used to build our models, and test sets - used to evaluate the prediction power of our models.

- **Goodness of fit** is defined as the extent to which the sample data are consistent with the model, which examines how well the model explains the data. MSPE (Mean Squared Prediction Error) on training sets is a measure of goodness of fit and a smaller MSPE indicates better goodness of fit.
- **Prediction power** measures how well models can predict future values. Mean squared prediction power is used to compare the prediction performance across all of our fitted linear predictive models. MSPE (Mean Squared Prediction Error) on test sets is used to measure the prediction power in this study and a smaller MSPE indicates better prediction power.

Refer to **Reference** for forumlas of MSPE.

Ordinary Linear models

The full linear model (*OLR full*) was built first. TaxClassCode, Municipalities, assessTotal, landTotal, improvementTotal and propertyCount were considered in this model. Based on the results of EDA, a list of significant variables were selected and included in another linear model (*OLR transformed*), as shown in **Table 2**. To compare the effect of linear models with and without features, a null model (*OLR null*) with no features used, was also constructed.

Refer to **Appendix** for variables selected in both *OLR full* and *OLR transformed* models.

Ridge and Lasso

Other than ordinary linear regressions, we are also interested in the performance of more advanced linear regressions like *Ridge* and *Lasso*. *Ridge* and *Lasso* have different objective functions to optimize; they take penalties in the sum of absolute values and sum of squared absolute values of weights respectively. A reason to consider Ridge Regression is that it helps deal with multicollinearity of the explanatory variables. This might be relevant to our study as some features used in this study are correlated (assessTotal, landTotal and improvementTotal). Lasso Regression is also included in this study because it does variable selections automatically by imposing a constraint on model parameters that may possibly cause some regression coefficients to shrink to zero.

Also, another advantage of these models is that they help reduce the variance of MSPE, so that MSPE is more stable. For more details about the two models, please refer to **Reference**.

Cross Validation

To examine the goodness of fit and prediction power, a 50-run of 10-fold cross-validation was performed in this study. For each run, a 10-fold cross-validation was used to train the five models and make predictions on training and test sets respectively. Then, the MSPEs calculated from the training sets and MSPEs calculated from the test sets were stored in vectors of corresponding models.

After the 50 runs, a vector of MSPEs on the training sets and a vector of MSPEs on the test sets for each model were therefore constructed successfully. Based on these vectors, side-by-side boxplots were used to show the mean and the spread of MSPEs on the training sets and test sets across all models respectively.

For more details about cross-validation, please refer to **Reference**.

Results

Exploratory Analysis

- Continuous Variables

Since the mill rate is mainly affected by the total assessment, which is the sum of total land assessment and total improvement assessment, scatter plots between mill rate and total assessment among different municipalities is created to show the relationship in **Figure 1**. Each color in the figure belongs to three tax classes of one municipality.

Scatter plots between mill rate and total land assessment, mill rate and total improvement assessment are similar to **Figure 1**. For more details about them, please refer to **Appendix**.

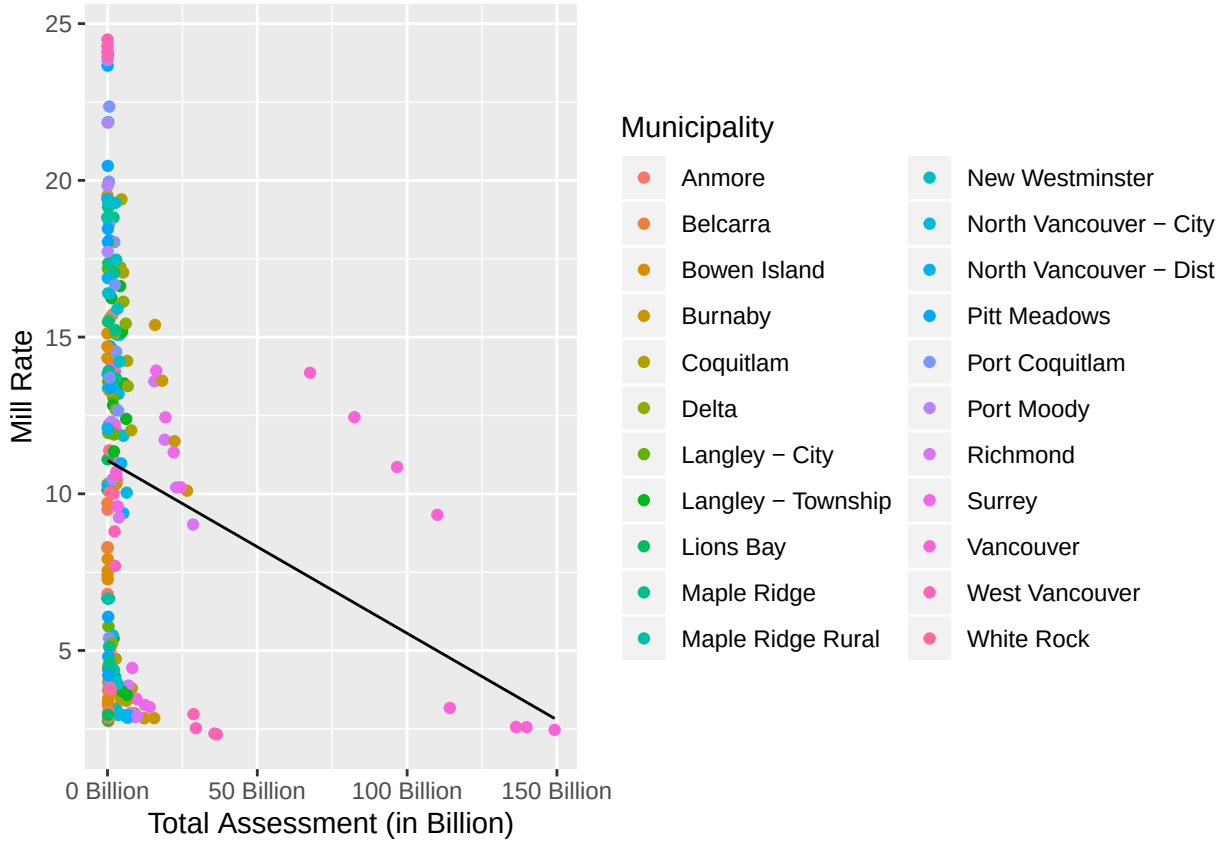

Figure 1: Mill Rate vs Total Assessment

There is no clear trend between the mill rate and total assessment from **Figure 1**. The plot has also shown that most points are condensed on the left horizontal axis since some municipalities have larger assessment values than others.

To reduce the effect of large assessment in some municipalities, total assessment across all municipalities is transformed by taking total assessment dividing by the number of properties of each municipality and tax class. The transformed data is named **"Average Total Assessment"**. A scatter plot between mill rates and average total assessments is shown in **Figure 2**.

Figure 2: Mill rate v.s. Average Total Assessment

The plot from **Figure 2** has shown that the mill rate tends to decrease as the average total assessment increase. Also, they have a moderately strong linear correlation.

- Categorical Variables

Here categorical variables are taken into account, boxplots of mill rate across municipalities and tax classes are plotted to display the distributions in **Figure 3** and **Figure 4**, respectively.

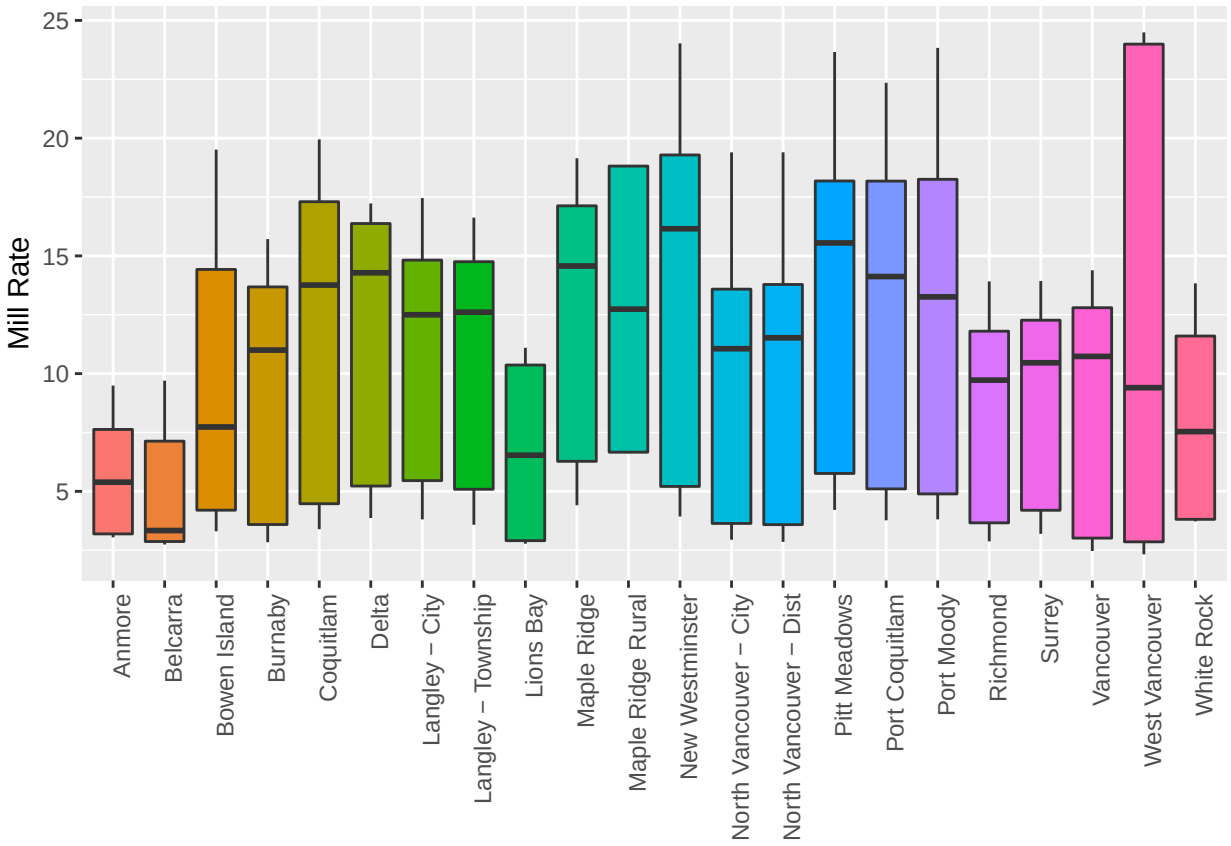

Figure 3: Mill rate across Municipalities

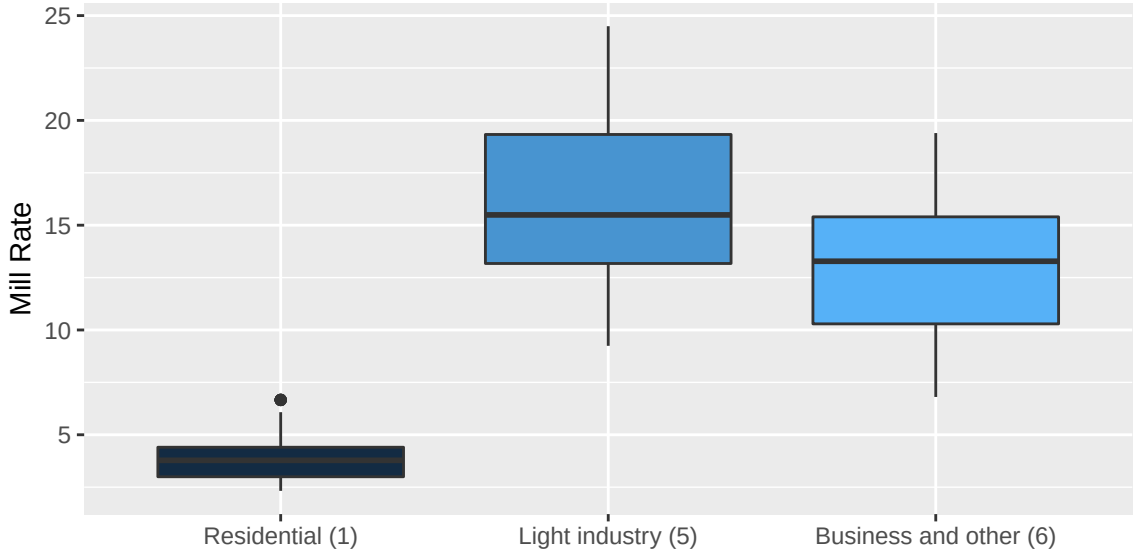

Figure 4: Mill rate across Tax Classes

Figure 3 has shown that most municipalities have different mean and variance in terms of mill rates. **Figure 4** has also supported that there is unequal mean and variance across tax classes.

A statistical test called the Kruskal-Wallis Test was performed to test their distribution. It was used to decide if population distributions were identical, and the corresponding p-value which is smaller than 0.05 indicated that the data have nonidentical distributions.

The results of Kruskal-Wallis Test is shown in **Table 1**.

Table 1: Kruskal-Wallis Test of Mill Rate across Municipality and Tax Class

Distribution	p-value
Mill Rate across Municipality	0.00675
Mill Rate across Tax Class	$<$ 2.2e-16

The p-values in **Table 1** have supported that there are nonidentical distributions of mill rates across municipalities and tax classes.

In conclusion, Average Total Assessment, Tax Class Code and Municipality are selected to fit another linear model, named **OLR transformed**. Refer to **Appendix** for more detail.

Linear Models

Below is a comparison of the five models (*OLR full*, *OLR transformed*, *Lasso*, *Ridge* and *OLR null*) using the goodness of fit and prediction power. The goodness of fit was measured by MSPE on the training sets, whereas prediction power was measured by MSPE on the test sets. Generally, smaller MSPEs on the training sets and test sets indicate better fit and prediction power respectively.

The distributions of the goodness of fit and prediction power across models are displayed in **Figure 5** and **Figure 6**, respectively, the distribution of null model is removed since it has larger values compared to all other models.

The distributions of MSPE on the training sets and test sets across all models are also displayed in **Table 2** and **Table 3** respectively.

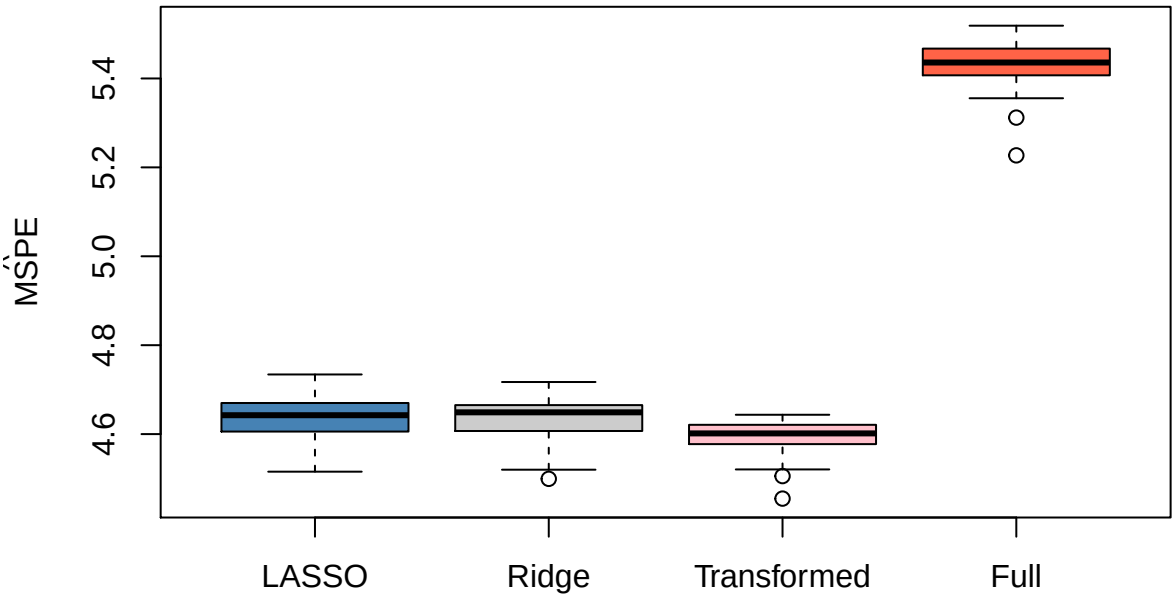

Figure 5: Goodness of Fit

Table 2: Goodness of fit of All Models

Model	PMSE:Min	PMSE:Mean	PMSE:Max
Full Model	5.23	5.43	5.52
Transformed Model	4.45	4.59	4.64
Null Model	36.33	36.52	36.64
Lasso for Full Model	4.52	4.63	4.73
Ridge Regression for Full Model	4.50	4.63	4.72

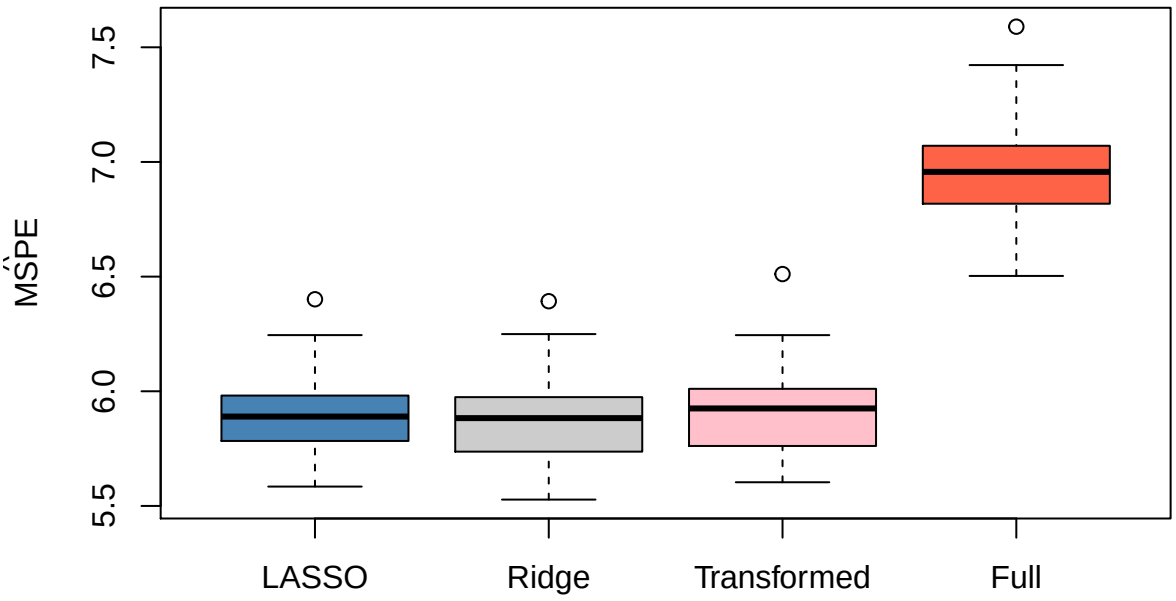

Figure 6: Prediction Power

Table 3: Prediction Power of All Models

Model	PMSE:Min	PMSE:Mean	PMSE:Max
Full Model	6.50	6.97	7.59
Transformed Model	5.60	5.92	6.51
Null Model	36.67	36.87	37.26
Lasso for Full Model	5.58	5.91	6.40
Ridge Regression for Full Model	5.53	5.88	6.39

As for goodness of fit, *Lasso*, *Ridge*, and *OLR transformed* have similar performance; MSPEs across these models were around 4.6, but *OLR transformed* performed slightly better and had a smaller spread (variance) of MSPE. *OLR full* performed worse than the other three, around 5.4, and *OLR null* had the greatest MSPE, around 35.

Similarly, as for prediction power, *Lasso*, *Ridge*, and *OLR transformed* performed roughly the same; MSPEs across these models were close to 6. *OLR full* was worse with MSPE around 6.9. *OLR Null* performed the worst with MSPE over 35.

Therefore, we conclude that from the results of a 50-run of 10-fold cross-validation, *Lasso*, *Ridge*, and *OLR transformed* had the best goodness of fit and prediction power. *OLR null* perfromed much worse than all the other four models.

Conclusion

From the exploratory data analysis, we have found that Total Assessments contain outliers in major municipalities such as Vancouver and Burnaby. In order to reduce the effect of outliers, the Total Assessment was transformed by taking its average over the number of properties in each municipality. The correlation analyses had shown that the transformed assessment total is the only continuous variable that has a relatively strong correlation with the mill rate. The Kruskal-Wallis had suggested that mill rates in each municipality and tax class are significantly different. From these results, we have decided to use Average Assessment Total, TaxClassCode and Municipality to fit our transformed model.

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 the client prefers a simpler model, we choose the *transformed model* to make our 2020 prediction.

For predicted mill rate in 2020, please refer to **Appendix Table 6**.

Appendix

Variables selected in different models

Table 4: Selected Explanatory Variables of OLR full

Parameter	Type
Total Assessment	Quantitative
Total Land Assessment	Quantitative
Total Improvement Assessment	Quantitative
Total Number of Properties	Quantitative
Tax Class Code	Categorical
Municipality	Categorical

Ridge Regression and *Lasso* models were fitted based on *OLR full* model.

Table 5: Selected Explanatory Variables of OLR transformed

Parameter	Type
Average Total Assessment	Quantitative
Tax Class Code	Categorical
Municipality	Categorical

Scatter plots:

- mill rate v.s. total land assessment

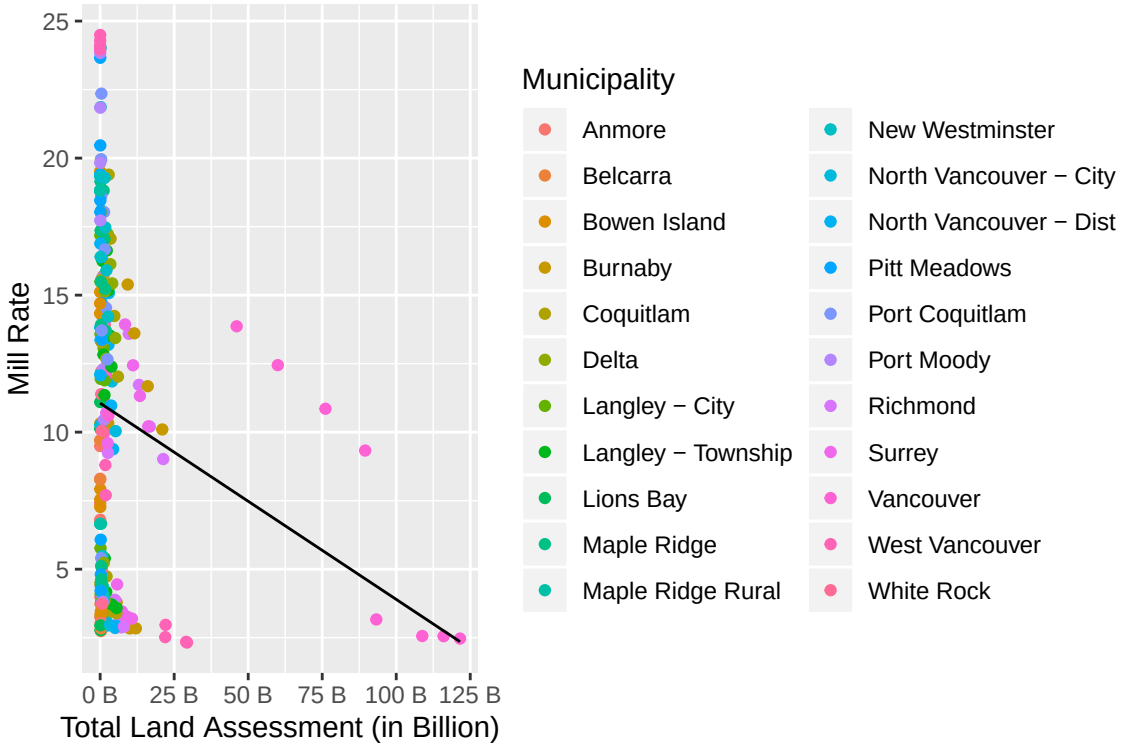

Figure 7: Mill Rate vs Total Land Assessment

- mill rate v.s. total improvement assessment

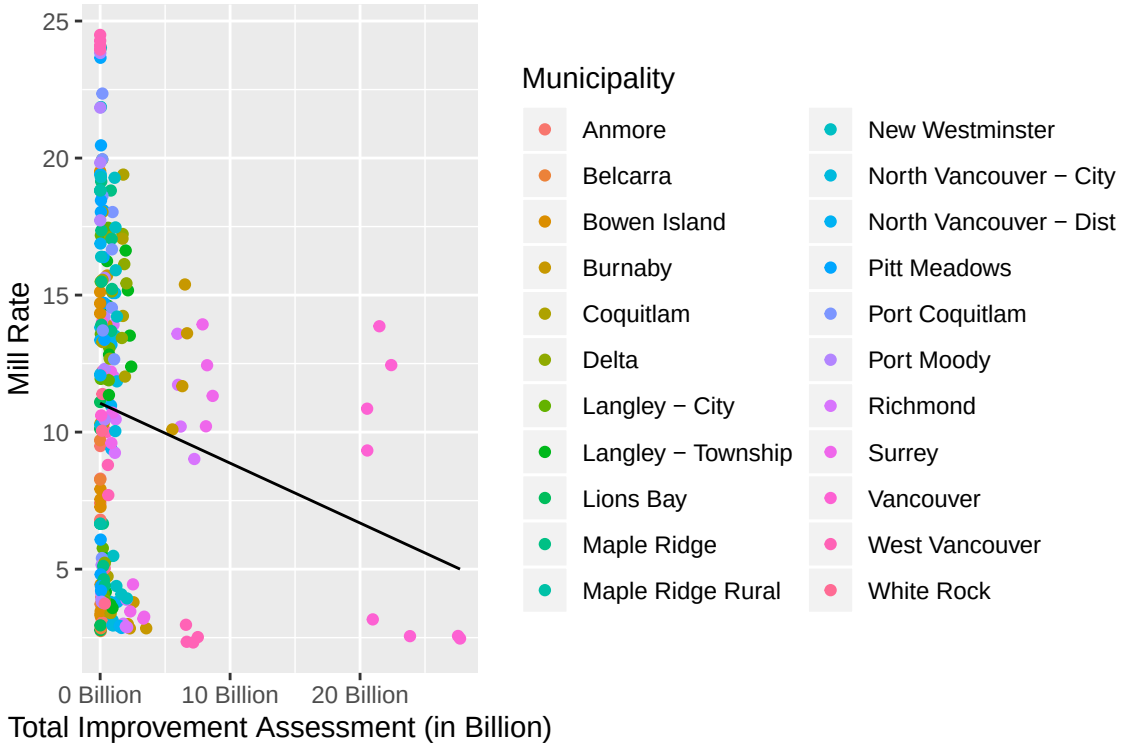

Figure 8: Mill Rate vs Total Improvement Assessment

2020 Mill Rate prediction

Table 6: Predicted Mill Rate in 2020 based on transformed model

Municipality	Residential (%)	Light industry (%)	Business (%)
Anmore	0.98	NA	9.91
Belcarra	1.19	NA	12.31
Bowen Island	1.63	14.69	11.15
Burnaby	-2.76	12.80	10.88
Coquitlam	1.72	17.70	12.84
Delta	5.50	12.98	13.65
Langley - City	1.73	15.12	12.25
Langley - Township	3.77	12.46	12.49
Lions Bay	2.19	NA	11.66
Maple Ridge	5.11	16.88	13.60
New Westminster	4.65	18.04	15.78
North Vancouver - City	-2.17	15.51	11.90
North Vancouver - Dist	1.29	13.68	11.82
Pitt Meadows	6.06	17.00	13.78
Port Coquitlam	4.91	17.58	14.50
Port Moody	4.16	17.66	13.44
Richmond	1.14	10.42	10.29
Surrey	0.82	12.47	10.53
Vancouver	2.69	12.38	8.79
West Vancouver	4.92	17.05	12.62

Municipality	Residential (%)	Light industry (%)	Business (%)
White Rock	1.88	16.33	12.17

Missing Value:

There are 1801 missing values in mill rate(TaxClassTaxRate). We decided to impute these missing values Based on client information, all properties in the same region, classcode, and year should have a unique class rate

- For entries with mill rate, we aggregated them into groups by region + classcode + year.
- For entries without mill rate, we found the group they belong to and assign them mill rate in that group.

Here is some exceptions found:

Some groups' mill rate is not unique:

- In Delta, properties in different neighbourhoods have slightly different mill rates. Since the variance is not significant, we take the mean as the overall mill rate in groups.
- In New Westminister, 2019, Class 06, one property's mill rate is different from others. It is regarded as an outlier.
- In Vancouver, 2019, Class 01, one property's mill rate is different from others. It is regarded as an outlier.
- In Burnaby, 2019, Class 06, six properties' mill rate are different from others. They are regarded as outliers.
- In Langley, 2019, Class 06, mill rate is different between assessment type. After talking to the client, the mill rate for assessment type "land" is regarded as the overall mill rate in that group.
- In some groups, all entries' mill rates are missing. Entries in these groups are removed. Here is the list of the groups:

Table 7: Data with missing mill rate

Year	Region	Class	Number of Properties
2016	Belcarra	06	9
2016	Lions Bay	01	40
2016	Lions Bay	06	25
2016	Maple Ridge Rural	05	36
2017	Belcarra	06	9
2017	Lions Bay	01	39
2017	Lions Bay	06	24
2017	Maple Ridge Rural	05	36
2018	Maple Ridge Rural	05	36
2019	Maple Ridge Rural	05	38

References

Code repository:

- Data Cleaning (https://github.com/STAT450-550/RealEstate/blob/450/src/Data_Cleaning.Rmd)
- Exploratory Data Analysis and Model Fitting (https://github.com/STAT450-550/RealEstate/blob/450/src/EDA%26mode_fitting.Rmd)

Measure of Goodness of Fit and Prediction Power: https://channabasavagola.github.io/2018-01-09-metrics/

Lasso and Ridge: https://web.stanford.edu/class/stats202/content/lec14-cond.pdf

Cross Validation: https://github.com/msalibian/STAT406/tree/master/Lecture2

Kruskal Wallis: http://www.biostathandbook.com/kruskalwallis.html