STAT 450 Project: Real Estate

Xuechun Lu, Yuting Wen, Peter Han, Yuetong Liu

15/03/2020

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 has shown there are strong relationships between mill rate and tax class and mill rate and municipalities; it has also shown 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 *the 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.

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

Data Description

Our client provides us the past 5 years' property assessment data in BC. Since the only interest is in predicting mill rate for metro Vancouver and specific tax class code, 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 is 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)
- Year

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 vs. Assessment, Mill Rate vs. Land Total, and Mill Rate vs. Improvement Total were performed. From our initial analysis, outliers in municipalities had been 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 vs. tax class and mill rate vs. municipality.

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

Assumption of Linear Regression

Models used in this study are mainly linear regressions. Before model construction, it is important to test the two assumptions of linear regression, i.e. normality of data and equal variance across groups for each categorical variables. If the two assumptions are satisfied, then the test results can be trusted.

A histogram of Mill Rate and Normal Q-Q Plot were used to test the normality assumption. If the data is normally distributed, the histogram would be symmetric with peak in the middle and decreasing frequency as approaching the two tails and the data in the Normal Q-Q Plot would mostly fall on a 45 degree line.

Fligner-Killeen Test was used to test the homogeneity of variances. The test does not rely on the assumption of normality and thus it is more robust to non-normal data. We applied the test to each categorical variable to test if there is a equal variance aross all sub-groups in each variable. If p-value is less than 5% for one categorical variable, then at 5% significane level, there is a significant difference of variance aross groups for that variable.

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. The definition of the goodness of fit and prediction power is given below.

- **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 was used to compare the prediction performance across all of our fitted linear predictive models. Prior to examine the prediction power, the data set was divided into training data - used to build our models, and testing data - used to evaluate the prediction power of our models. MSPE (Mean Squared Predction Error) on testing sets is used to measure the prediction power in this study.

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

Ordinary Linear models

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

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 a different objective function to optimize; they take a penalty 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. Lasso Regression is also included in this study because it does variable selections automatically by imposing a constraint on model parameters that causes some regression coefficients to shrink to zero.

Also, another advantage of these models is that MSPE is more stable as the variance of MSPE is reduced. For more detials 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 testing data respectively. Then, the MSEs calculated from the training data and MSPEs calculated from the testing data were stored in vectors of corresponding models.

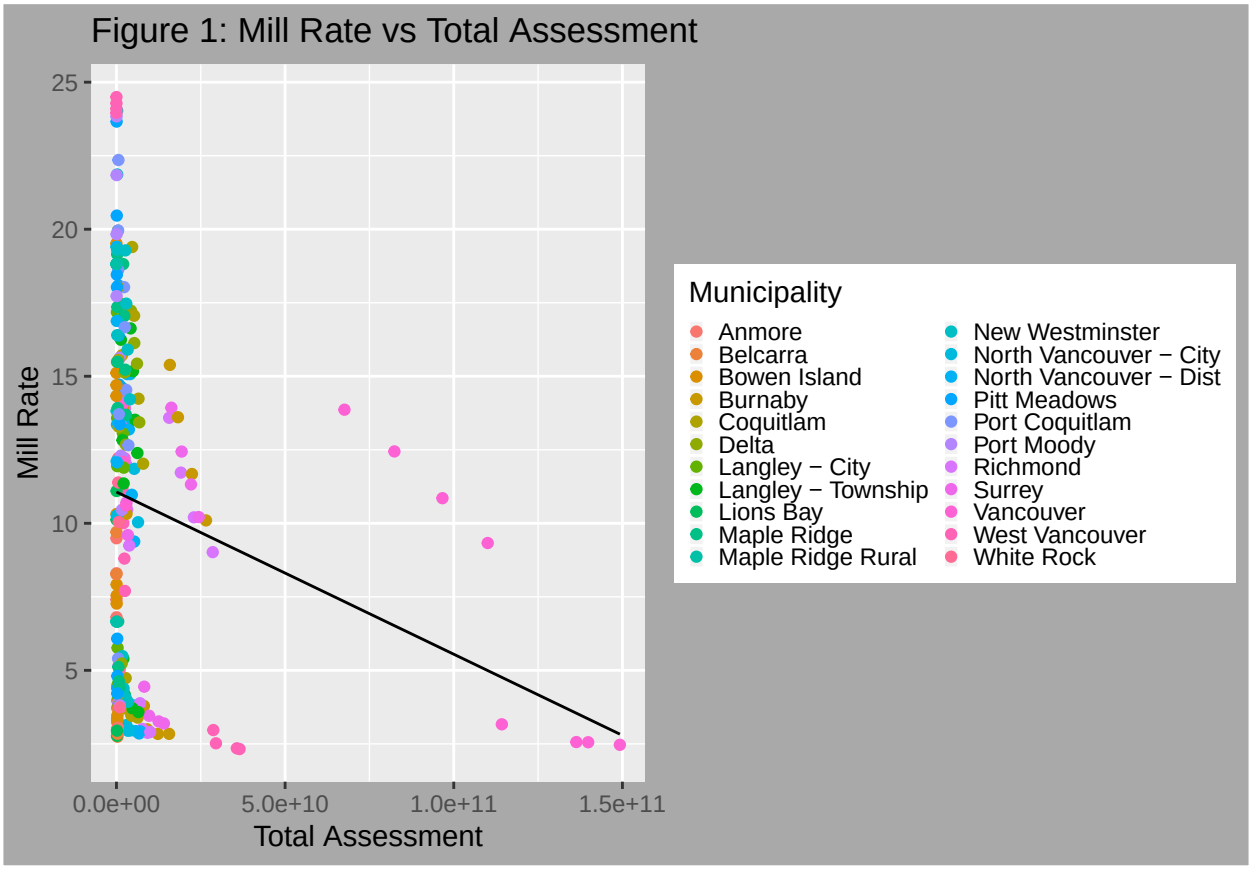
After the 50 runs, a vector of MSEs and a vector of MSPEs for each model were therefore constructed successfully. Based on these vectors, side-by-side boxplots were used to show the mean and the spread of MSE and MSPE across all models respectively. For more details about cross-validation, please refer to **Reference**.

Results

Exploratory Analysis

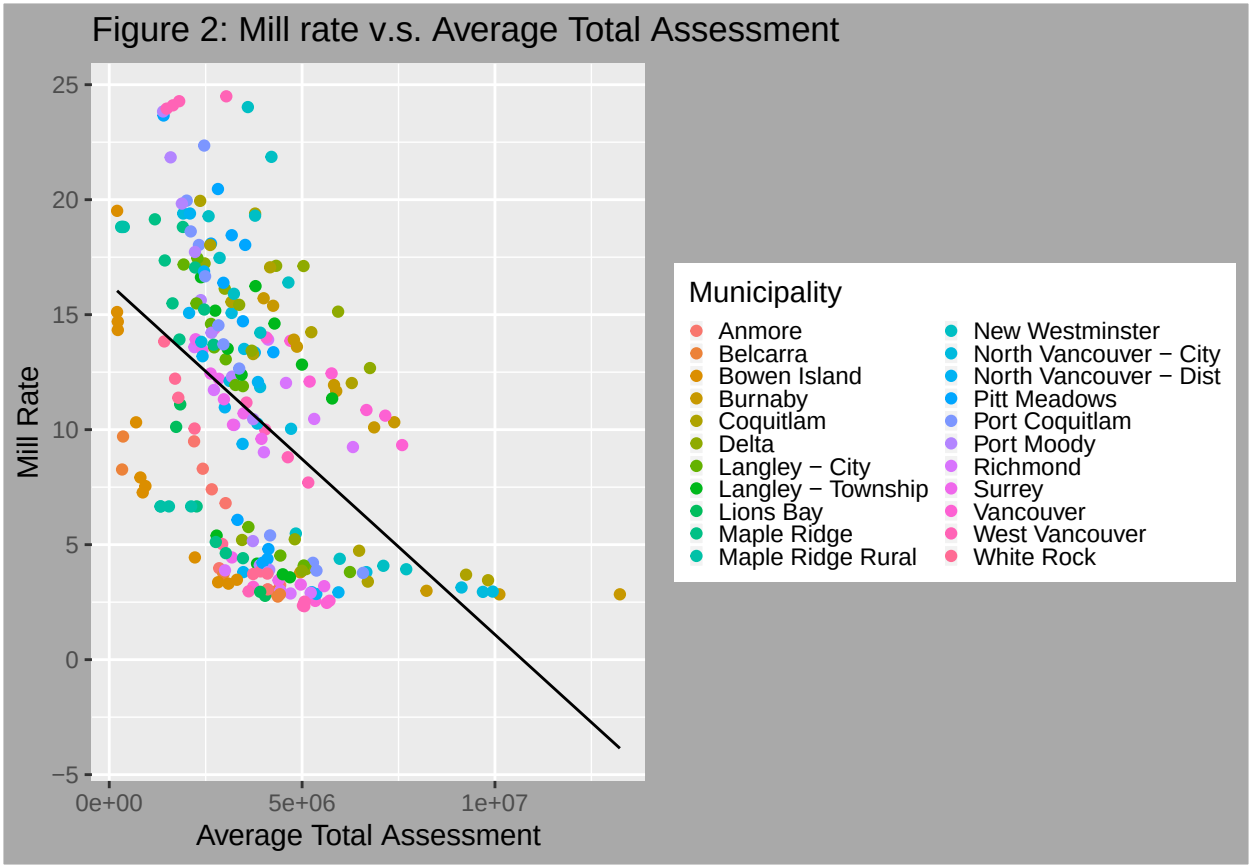
- Continuous Variables Since the mill rate is mainly affected by the total assessment, which is the sum of total land assessment and total improvement assessment, scatter plots between mill rate and total assessment among different municipalities is created to show the relationship in Figure 1. Each 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**.



There is no clear trend between the mill rate and total assessment. Plot has 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.



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.

Figure 3: Mill rate across Municipalities

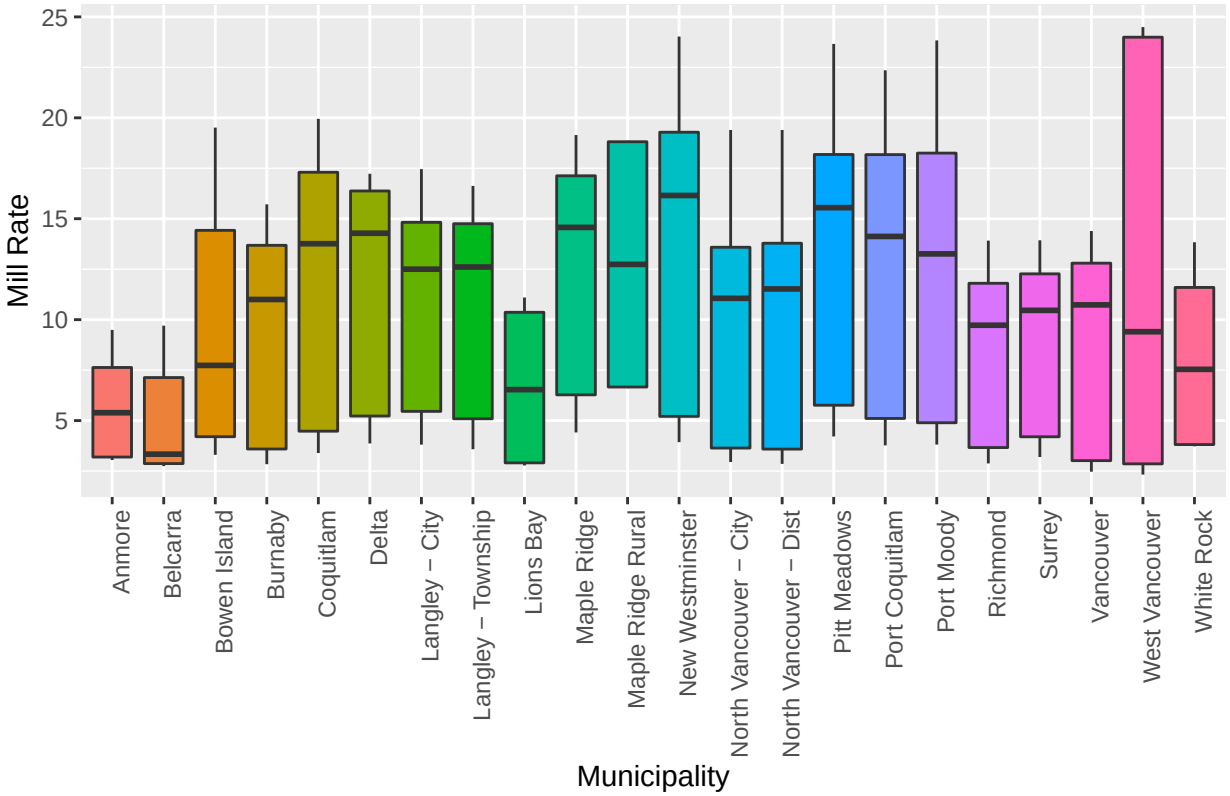


Figure 4: Mill rate across Tax Classes

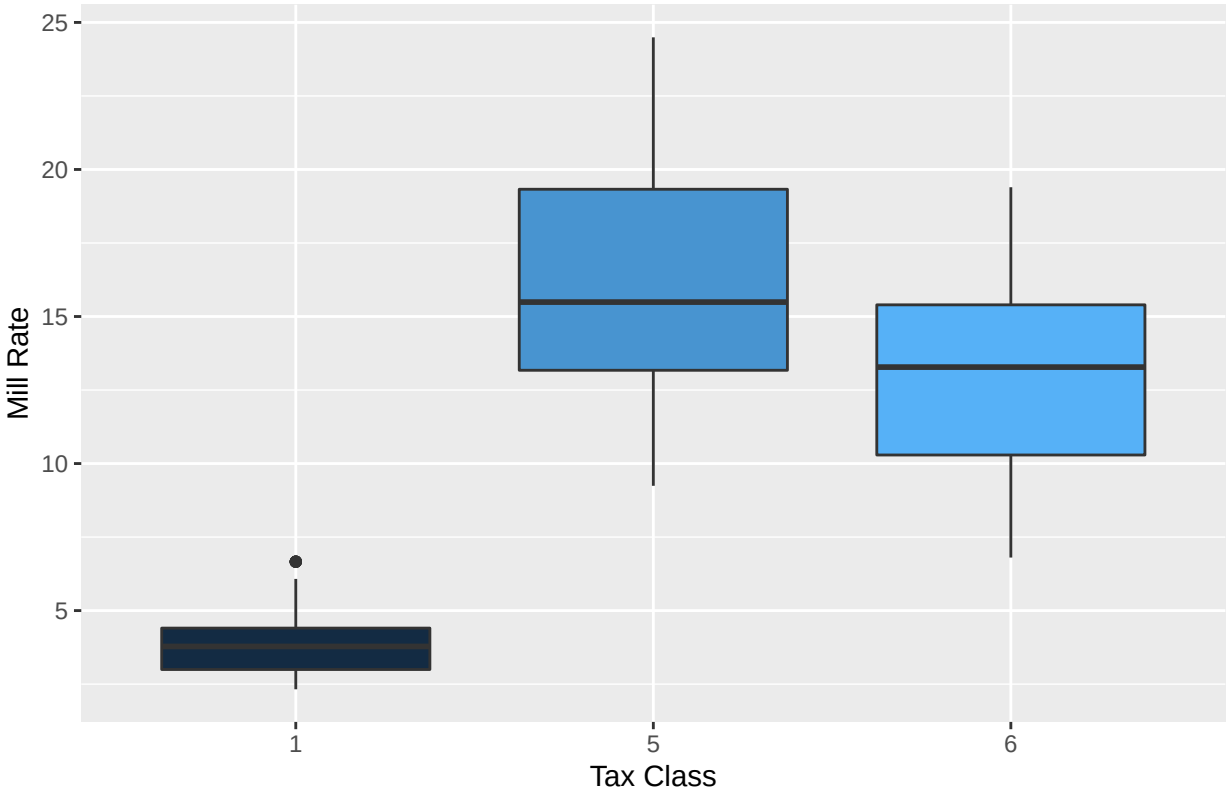


Figure 3 shows that most municipalities have different mean and variance in terms of mill rates.

Figure 4 also supports 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`

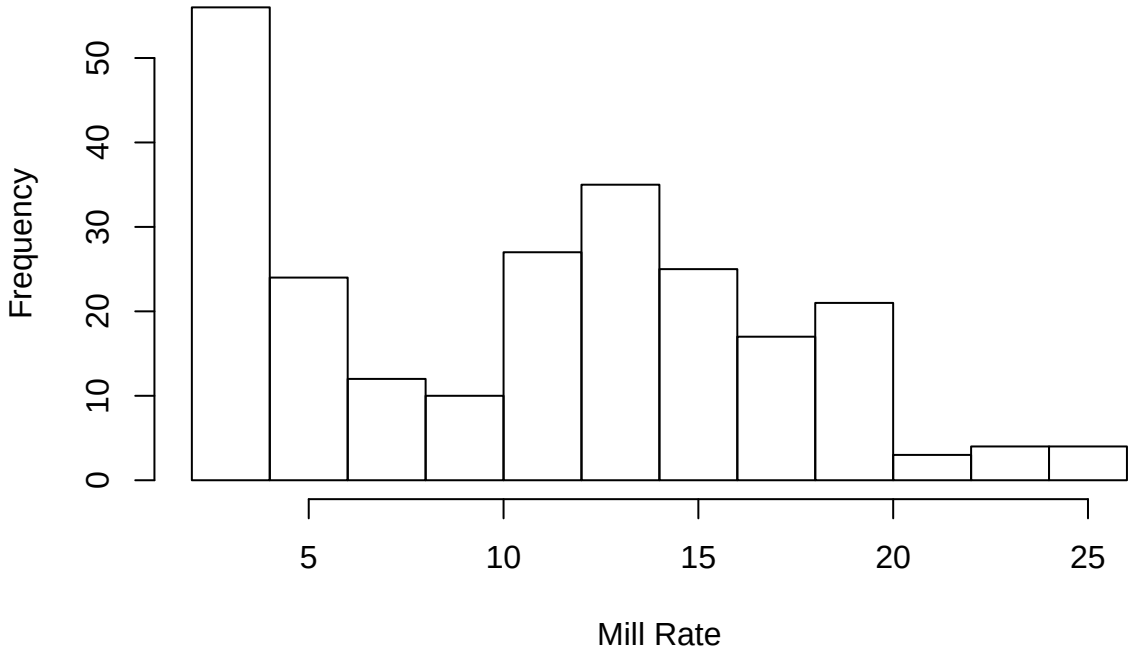
Distribution	p-value
Mill Rate across Municipality	0.00675
Mill Rate across Tax Class	$<$ 2.2e-16

The p-values in Table 1 support that there is nonidentical distributions of mill rates across municipalities and tax classes.

Assumptions of Linear Regression

1. Normality Assumption

Figure 5: Histogram of rate



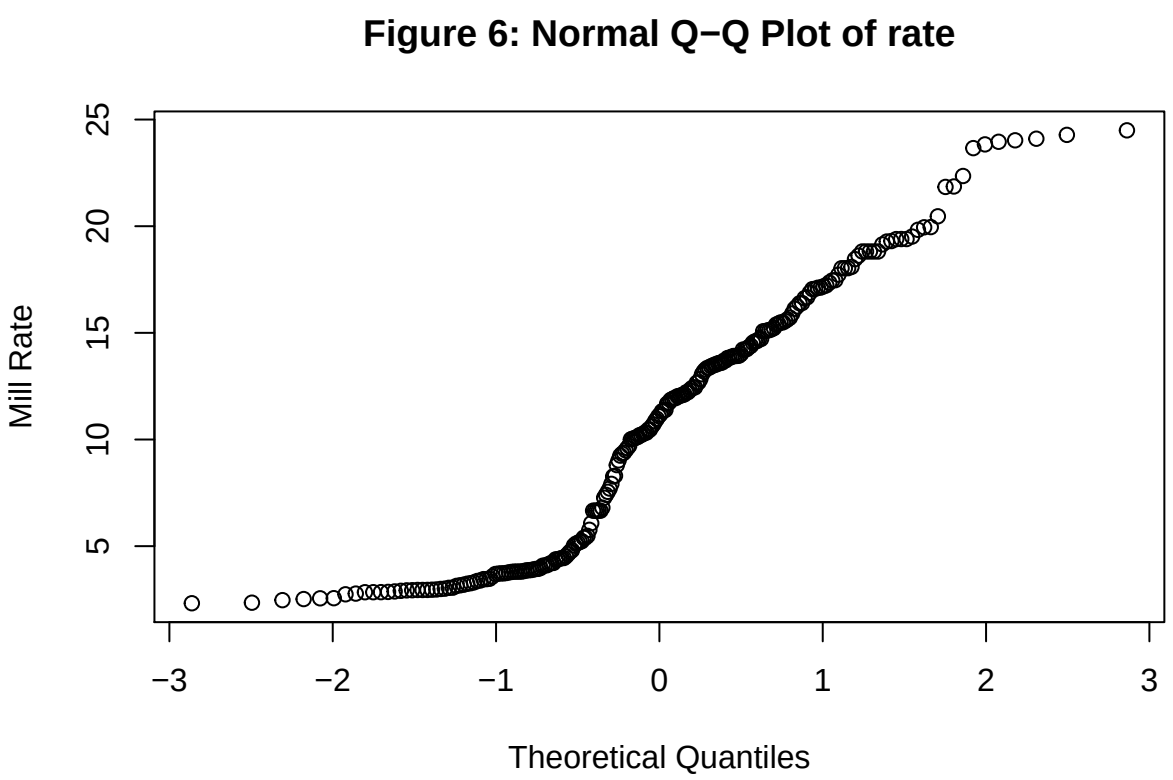


Figure 5 exhibits the distribution of mill rates and it is observed that rate is not normal and it has heavy tails on both ends, especially on the left hand side.

Figure 6 shows that most of data do not fall on the 45 degree line and curve away from the line at each end in opposite directions due to the "heavy tails" at the each end of the histogram. This, again, verifies that rate is not normally distributed.

2. Equal variance

Table 2: Fligner Killeen Test of Mill Rate across Municipality, Tax Class and Year

Distribution	p-value
Mill Rate across Municipality	0.1299
Mill Rate across Tax Class	< 2.2e-16
Mill Rate across Year	0.004907

Table 2 shows a non-significant difference between the group variances of Municipality, but a significant difference of variances in Year and Tax Class since p-values for Tax Class and Year are less than 5%. Therefore, the equal variance assumption is not satisfied for Tax Class and Year. This finding is also consistent with Figure 4 where it is clear that variance (spread) is not identical across all three tax classes.

Hence, assumptions of linear regression are not met in this study. This means the test results, for example, t-statisitcs and p-values, are biased and thus cannot be trusted. Since this study only focuses on prediction accuracy, the failure of linear regression assumptions might not matter much.

Linear Model Below is a comparison of the five models (*Lasso*, *Ridge*, *OLR transformed*, *OLR full* and *OLR null*) using the goodness of fit and prediction power. The goodness of fit was measured by MSE of the training data, whereas prediction power was measured by MSPE of the testing data. Generally, smaller MSEs and MSPEs 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 across models are also displayed in Table 3.

Figure 7: Goodness of Fit

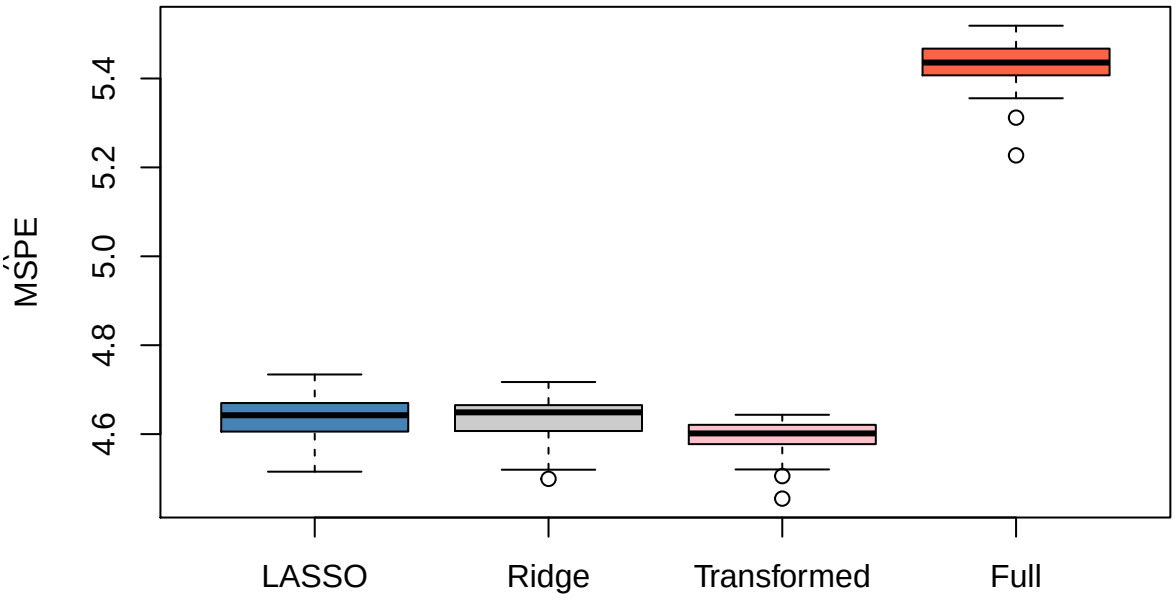


Figure 8: Prediction Power

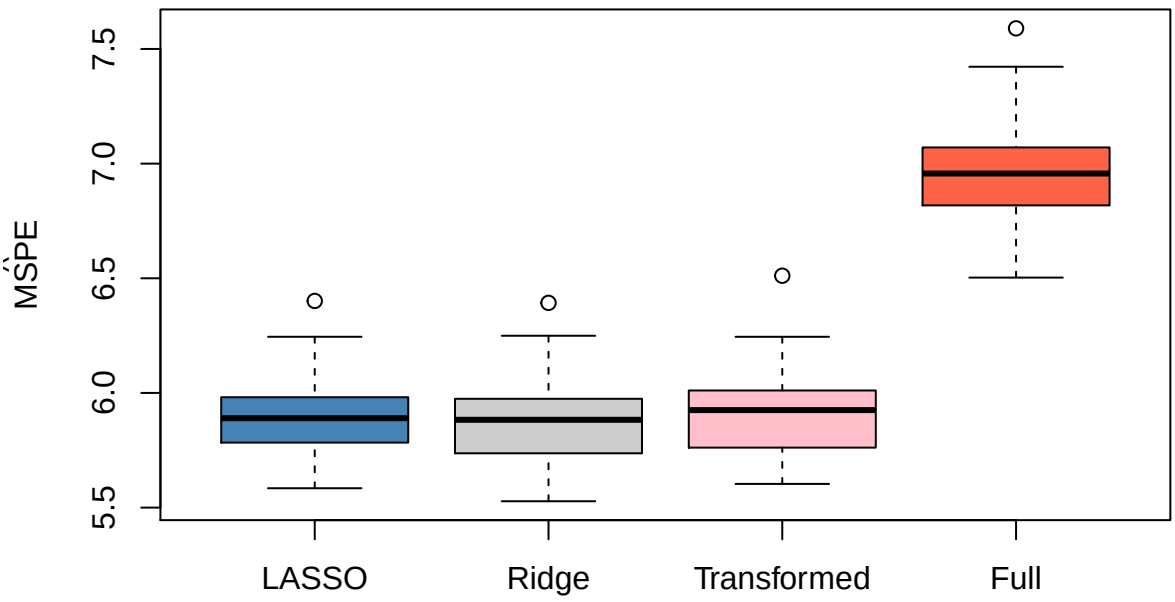


Table 3: Prediction Power of All Models

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
The 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 Assessments

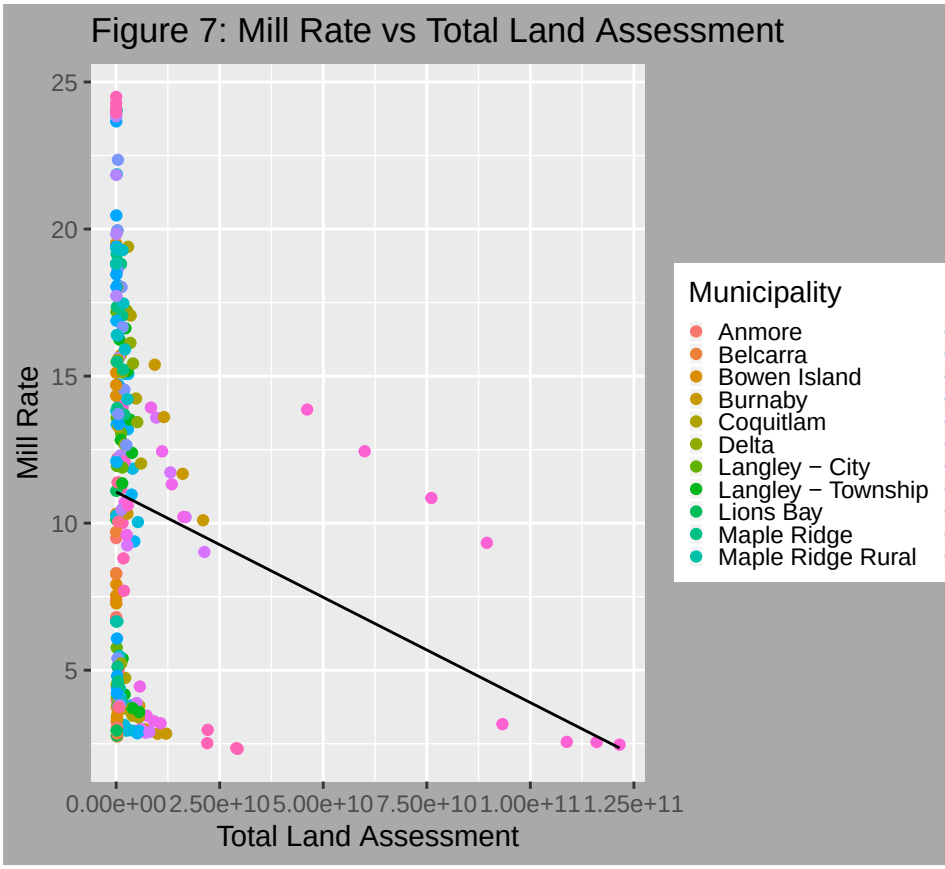
were transformed by taking their average over the number of properties in each municipality. The correlation analyses had shown that the transformed assessment total is the only continuous variable that has a relatively strong correlation with the mill rate. The Kruskal-Wallis had suggested that mill 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 reduced model.

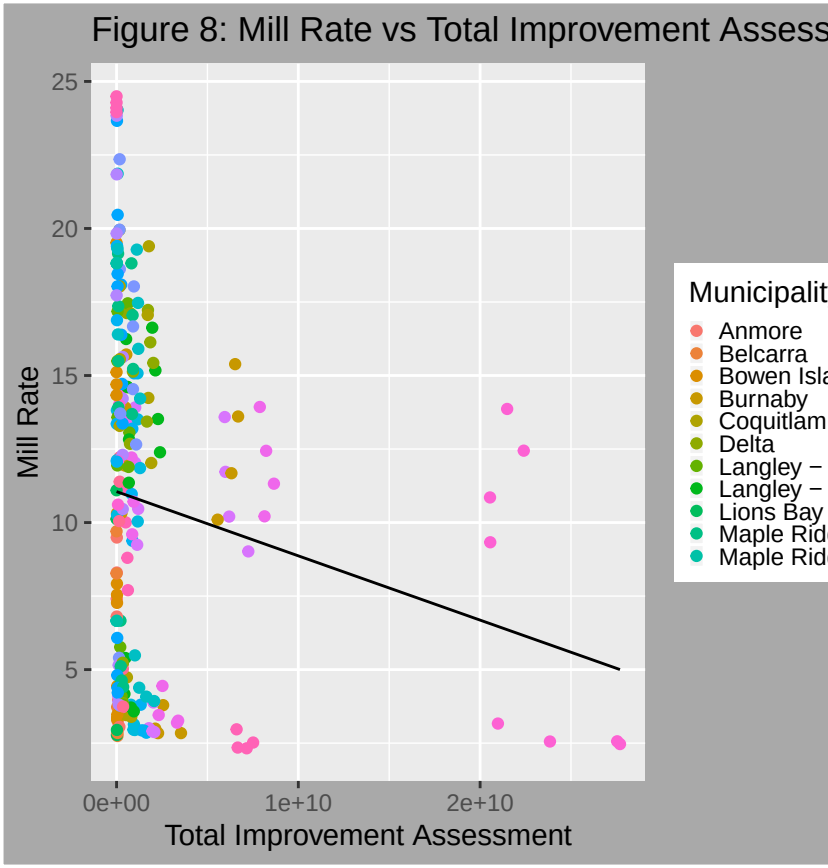
Transformed OLR, *Ridge Regression*, and *the LASSO* were able to make good predictions based on mean squared error and mean squared prediction error from cross-validation. Since the client prefers a simpler model, we choose the *transformed model* to make our 2020 prediction.

Appendix

Scatter plots:

- **mill rate v.s. total land assessment**





- mill rate v.s. total improvement assessment

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 3: 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