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 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 the following 22 municipalities:

- 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.

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 we are only interested in predicting mill rate for metro Vancouver and specific class code, we get a subset of properties that satisfy our interest:

- TaxClass Code in (01,05,06)
- Municipality in Metro Vancouver

Moreover, we select 5 features that could be relevant to mill rate, which are:

- 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, we aggregate all properties in the same region, tax class code, and year are aggregate into a group, becasue these properties have same mill rate, which is our response variable. Here is the summary statistics for these groups:

- Mill rate
- Total Assessment
- Total Land Assessment
- Total Improvement Assessment
- Total number of properties
- Taxclasscode
- Municipality
- 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 and found relationships 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 is 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.

Measure of goodness of fit and prediction power

In this study, we built linear models and evaluated their performances by 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 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. R squared and adjusted R squared are the most well known measures of goodness of fit and higher R squared and adjusted R squares indicate better goodness of fit.

- **Prediction power** measures how well models can predict the future values. We use mean squared prediction power to compare the prediction performance across all of our fitted linear predictive models. To do that, we will divide the data set into training data - used to build our models, and testing data - used to evaluate the prediction power of our models.

Ordinary Linear models

TaxClassCode, Municipalities, Assess Total, Land Assessment Total, Improvement Assessment total, Number of properties are considered. The full linear model (ORL full) using all the available features was built first. 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 penalty in sum of absolute values and sum of squared absolute values of weights respectively. These models have an interesting characteristic, that is, a weight is assigned to each feature. This might be relevant to our study as mill rate can be affected by each feature to a different extent. Also, another advantage of these model is that MSPE is more stable as the variance of MSPE is reduced.

Cross Validation

To examine the goodness of fit and prediciton power, a 50-run of 10-fold cross validation would be performed in this study. For each run, a 10-fold corss 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 mdoels.

After the 50 runs, a vector of MSEs and a vectors of MSPEs for each model were therefore constructed successfully. Based on these vectors, boxplots would be used to show the mean and spread of MSE and MSPE aross all models respectively.

Results

Exploratory Analysis

Continuous Variables

Since mill rate is mainly affected by total assessment, which is sum of total land assessment and total improvement assessment, scatter plots between mill rate and all kinds of assessment among different municipalities are created to show the relationship in Figure 1, Figure 2, Figure 3

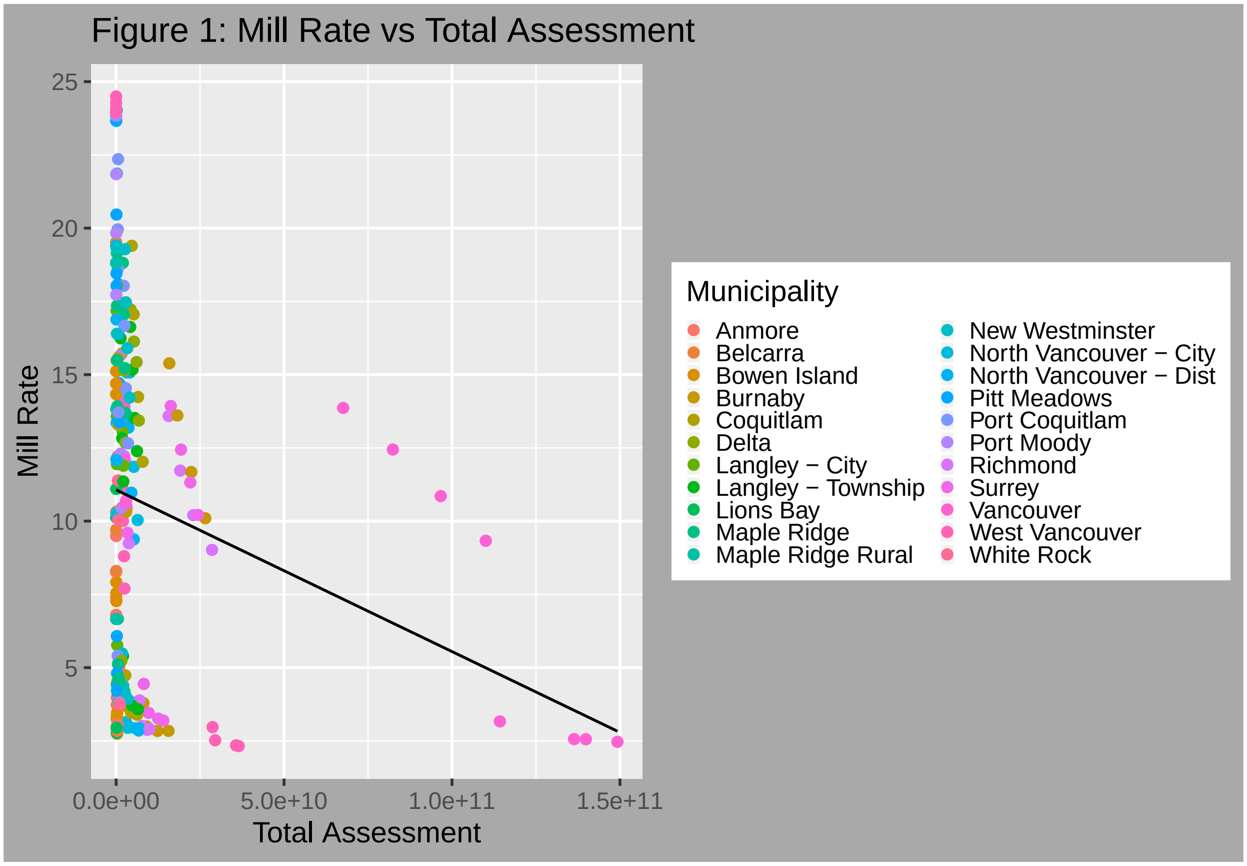


Figure 1: Mill Rate vs Total Assessment

Figure 2: Mill Rate vs Total Land Assessment

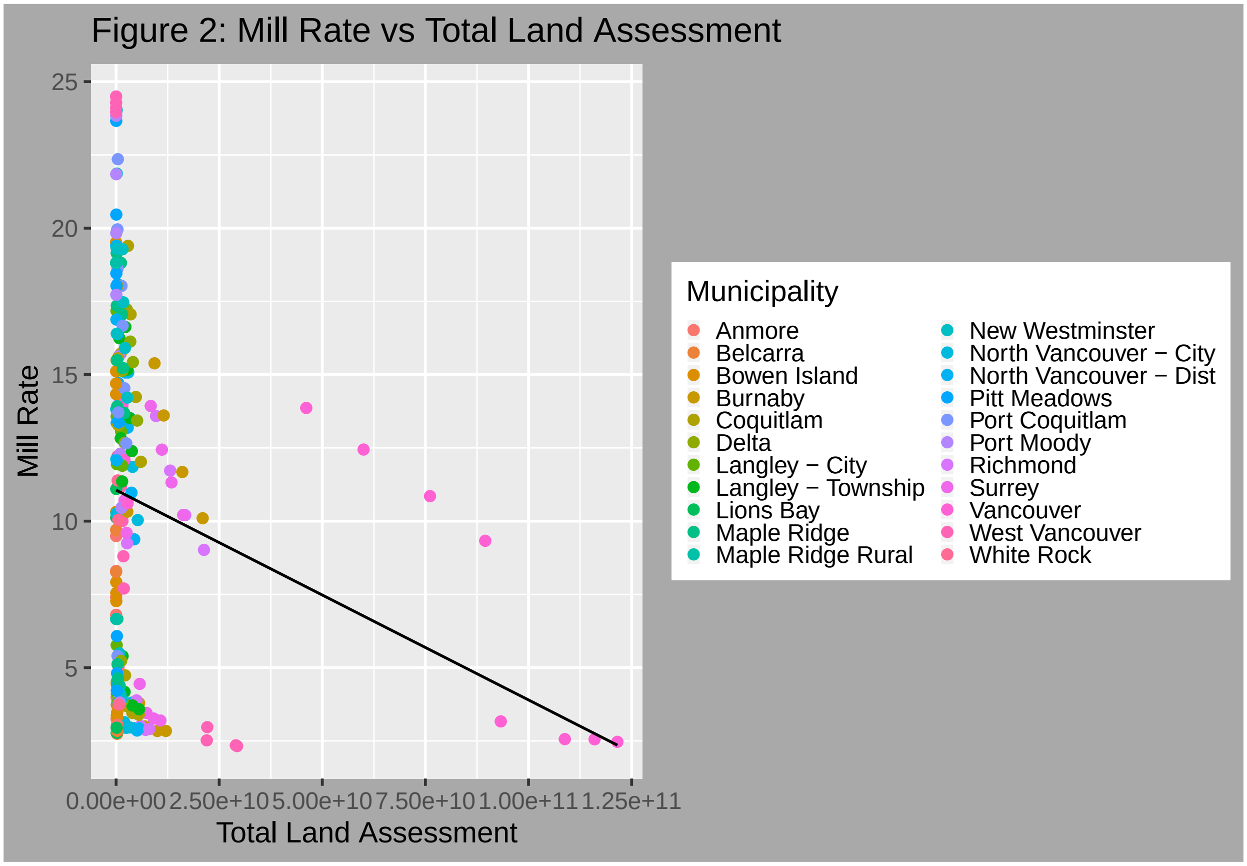
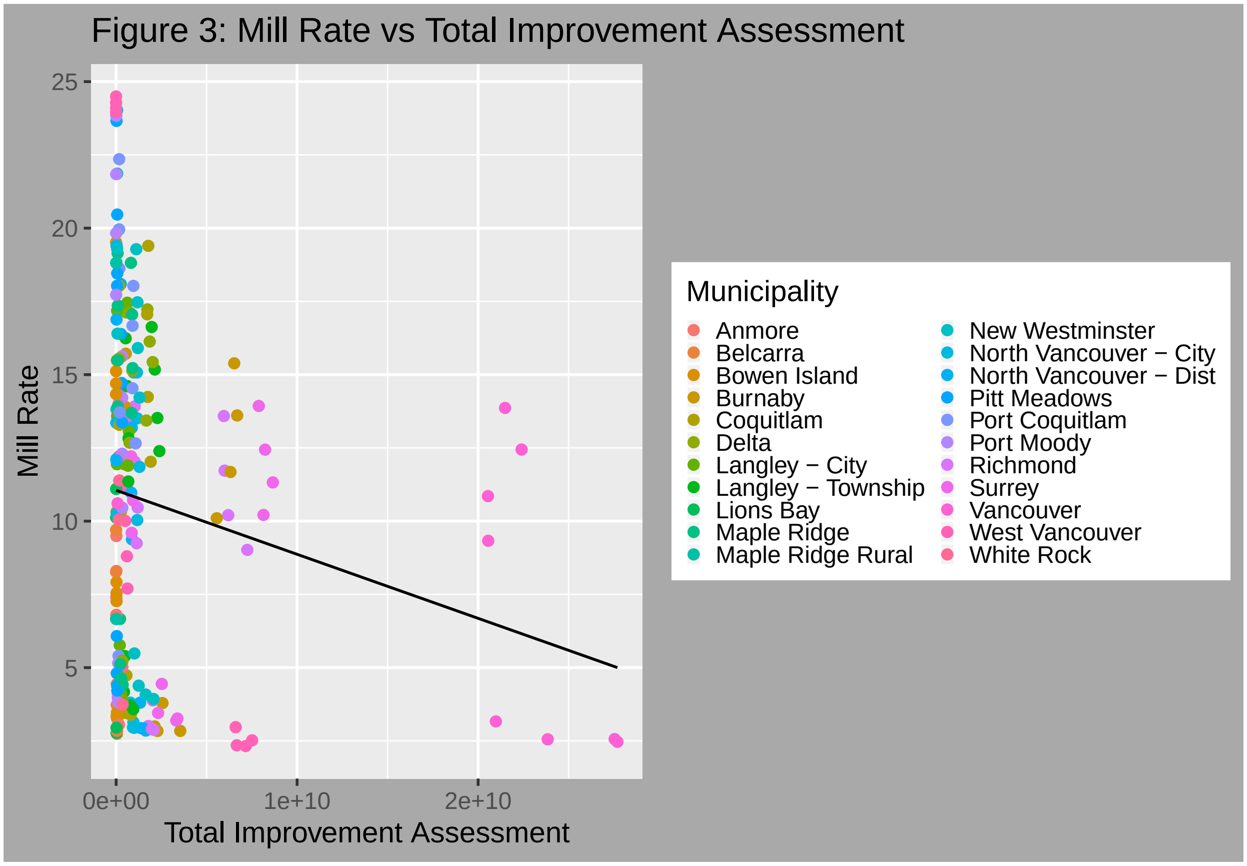
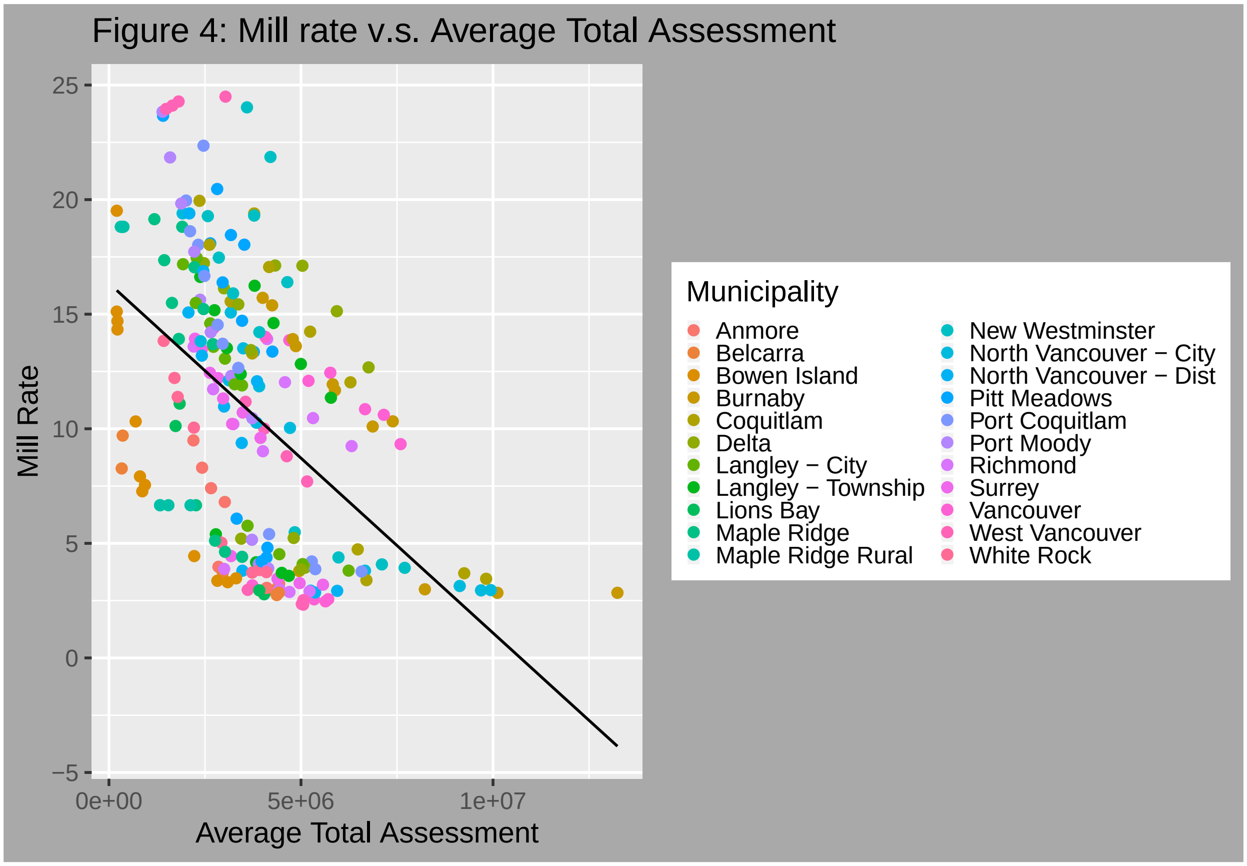


Figure 3: Mill Rate vs Total Improvement Assessment



There is no clear trend between mill rate and any kind of assessment. Plots also have shown that most points are condensed on the left horizontal axis, since some municipalities have large assessment.

To reduce the effect of large assessment of some municipalities, total assessment across municipalities are transformed by taking total assessment dividing by number of properties of each municipality and tax class. The transformed data is named **"Average Total Assessment"**. A scatter plot between mill rate and average total assessment is showed in Figure 4.



Plot from Figure 4 has shown that mill rate tends to decrease as average total assessment increase. Also they have moderately strong linear correlation.

Categorical Variables

Here categorical variables were taken into account, boxplots of mill rate across municipalities and tax classes are plotted to display the distributions in Figure 5 and Figure 6.

Figure 5: Mill rate across Municipalities

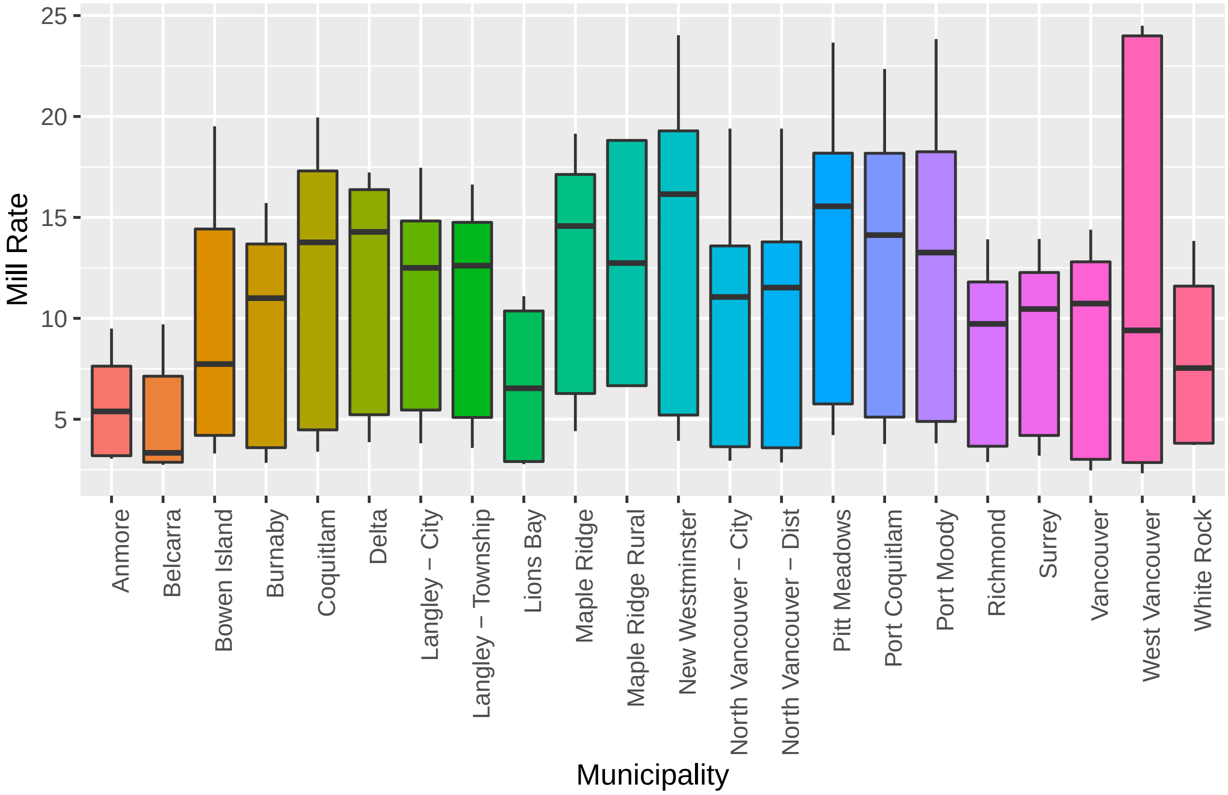


Figure 6: Mill rate across Tax Classes

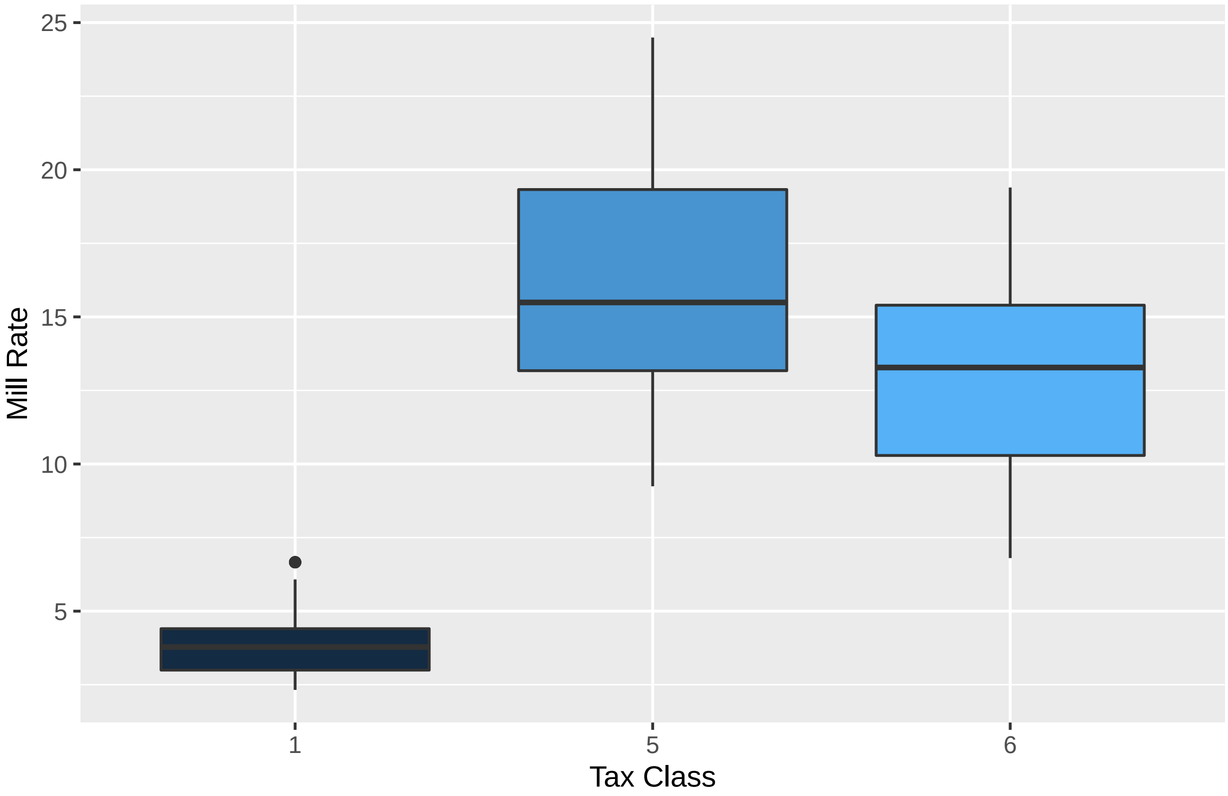


Figure 5 has shown that most municipalities have different mean and variance in terms of mill rates.

Figure 6 also supports that there is unequal mean and variance across tax classes.

A statistical test called Kruskal-Wallis Test is also performed to test their distribution. It is used to decide if population distributions are identical, and the corresponding p-value which is smaller than 0.05 can indicate that the data has 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	$< 2.2e\text{-}16$

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

Linear Model

Below is a comparison of the five models (Lasso, Ridge, OLR transformed, OLR full and OLR null) as for goodness of fit and predcition power. Goodness of fit is measured by MSE of the training data, whereas prediction power is measured by MSPE of the testing data. Generally, smaller MSEs and MSPEs indicate better fit and prediction power respectively.

The distributions of goodness of fit and prediction power across models are displyed in Figure 7 and Figure 8, 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 2.

Figure 7: Goodness of Fit

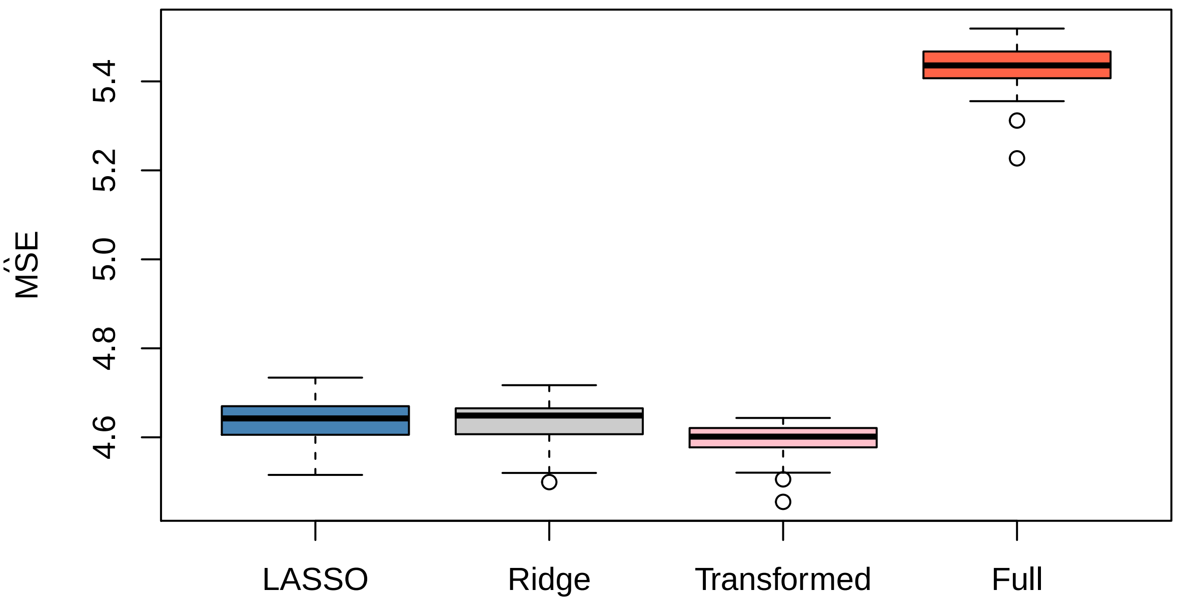


Figure 8: Prediction Power

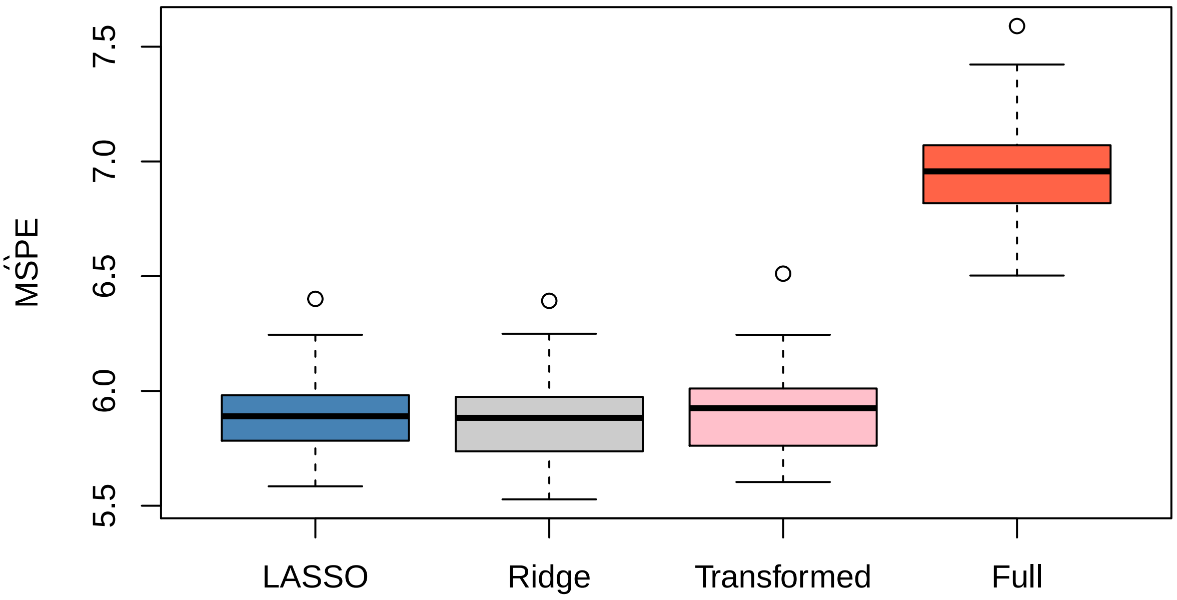


Table 2: Prediction Power of All Models

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 reduced have a similar performance; MSEs are around 4.6, but OLR transformed perform slightly better and has a smaller spread (variance) of MSE. OLR full performs worse than the the three, around 5.4, and OLR null has the greatest MSE, around 35.

Similarly, as for prediction power, Lasso, Ridge and OLR transformed perform roughly the same; MSPEs are close to 6. OLR full is worse with MSPE around 6.9. OLR Null performs 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 have the best goodness of fit and prediction power. OLR null perfroms much worse than all the other four models.

For a detailed look of how the models were fitted and evaluated, please refer to the **Appendix**.

TODO: delete pct change part, add prediction based on our chosen model (Peter)

Conclusion

From our exploratory data analysis, we can see that assessment total, land total, tax code, year and municipalities are all significant in our model. The client suggested that we can transform numerical factors such as assessment total into percentage change. This data transformation does not perform as well as we expected, which only yields an Rˆ2 of around 0.2 across all of our fitted models. We believe the method that the client has suggested might leave out some important information about the housing market in each municipality. We have also found that when comparing the correlation between mill rate against features in our models, Vancouver is an outlier. We would like to further investigate Vancouver as a special case. We are also interested in the effect of different tax classes in predicting mill rate. Further analysis and prediction will be performed based on our hypothesis.

All fitted linear model were able to make accurate prediction based on means squared error, but there is still a very high chance that our model is overfitted. For our next report, we are going to use cross-validation to reduce the effect of overfitting. ##____

Appendix

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)