

STAT 450 Project: Real Estate

Yuetong Liu, 43838168

3/2/2020

Summary

Our client is a real estate data analyst in Metro Vancouver who analyzes the market trend to make a market prediction for his company and future client. He is interested in accurately predicting future mill rates in Metro Vancouver. There are three major sections of the project: Data cleaning, exploratory data analysis, and linear modelling. Data cleaning aggregated our data into summary statistics and also reduced the number of data entries. Exploratory data analysis analyzed correlation between mill rate and other features in assessment data frame. Linear modelling gave the best goodness of fit and prediction power among reduced linear regression models, Ridge, Lasso and Elastic net.

Introduction

Our goal is to predict the property tax in 2020, for each of the following 21 sub-regions in the Greater Vancouver area. A full linear model, a reduced linear model, LASSO, Ridge, and elastic net Regressions were used to train our predictive model. Correlations between mill rate, assessment value, and municipal budget were performed to see the significance of these factors in our predictive model.

Once we have finished the above tasks, we will use the mean square test error to evaluate the performance of our model.

Data Description

Since we are interested in predicting mill rate, only the assessment data set provided by the client was used. There are approximately 2.4 million data entries across these 5 years.

The dataset was filtered based on the following conditions:

TaxClassCode: 01, 05, 06

MunicipalityDesc:

- Burnaby
- Coquitlam
- Delta
- Langley - City
- Langley - Township
- Maple Ridge
- Maple Ridge Rural
- 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

Year: 2016, 2017, 2018, 2019

Mill Rate and 4 features relevant to Mill Rate are selected:

- Mill Rates
- Assessment Value
- Tax Class Code
- Area Code
- Year

We also collected external data from the Government of British Columbia which could be associated with the tax rate:

- Municipal Budget of Cities in Metro Vancouver (Taxes Imposed & Collected, Schedule 706)

Based on client information, all properties in the same region, classcode, and year should have a unique class rate. Therefore, all data were grouped by **region + classcode + year**. The assessment value and count of each properties within a group was summed up. Moreover, we decided to calculate the growth rate of TotalAssessedValue, LandAssessedValue, ImprovementAssessedValue, Budget and Tax Rate for each municipality and their corresponding tax class.

Methods

Exploratory Analysis

Before we made any prediction on the future mill rate of Metro Vancouver's real estate market, we did some exploratory analysis to visualize the main characteristics of our dataset and found relationships between mill rate, assessment and municipal budget.

- It's assumed that the government aims to match its budget and its income by adjusting the mill rates; therefore, we plotted a graph of tax income for three different tax classes (01, 05, and 06) vs. budget through time to see the trend of tax income in different municipalities to make sure our assumption about the data is correct.
- Correlation analysis between mill rate and other features. We used scatter plot for numerical variables and ANOVA table for categorical variables.

Measure of goodness of fit and prediction power

In this study, we first built various models and we 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.

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. In this study, we use mean squared prediction power to compare the prediction performance across all the models. To do that, we will divide the data set into training data, used as building models, and testing data, used as computing mean squared prediction power.

Linear model

The explanatory variables we are interested in are Year, TaxClassCode, Municipalities, assessTotal_pct and tax (budget). We first explore the full linear model using all the available features. Since there are potential outliers in the data set, it might be helpful to refit the model without the outliers to check if the fit of the model is improved.

Selection of variables

A list of significant variables are included in the result of the full model. In the next step, we will refit the model with only the significant variables. Goodness of fit will decline with fewer features, however, the performance of prediction power might be improved.

Ridge, Lasso and Elastic Net

Except for ordinary linear regression, we are also interested in the performance of Ridge, Lasso, and Elastic Net models as they are also in the linear regression family, but more advanced than ordinary linear regression. In these two models, weights are assigned to features. Ridge regression and Lasso take penalty in sum of absolute values and sum of squared absolute values of weights respectively. Elastic net is a combination of Ridge and Lasso.

We build the three models using all features and compare their goodness of fit, and then will examine their prediction power on testing data.

Neural Network

In the current report, we are mainly focusing on predicting the mill rate using Linear Model. In the future, we will discuss with our supervisor and try to fit a neural network model if possible (R package: neuralnet or keras). Neural Network has the ability to learn and model non-linear and complex relationships. It is flexible and can be used for both classification and prediction. Since we have a large number of data, it makes Neural Network a very good candidate for our prediction model.

Report on the performance of the Neural Network will be included in the next report.

Results: Exploratory Analysis

Figure 1: Budget and Tax Income v.s. Year of Vancouver

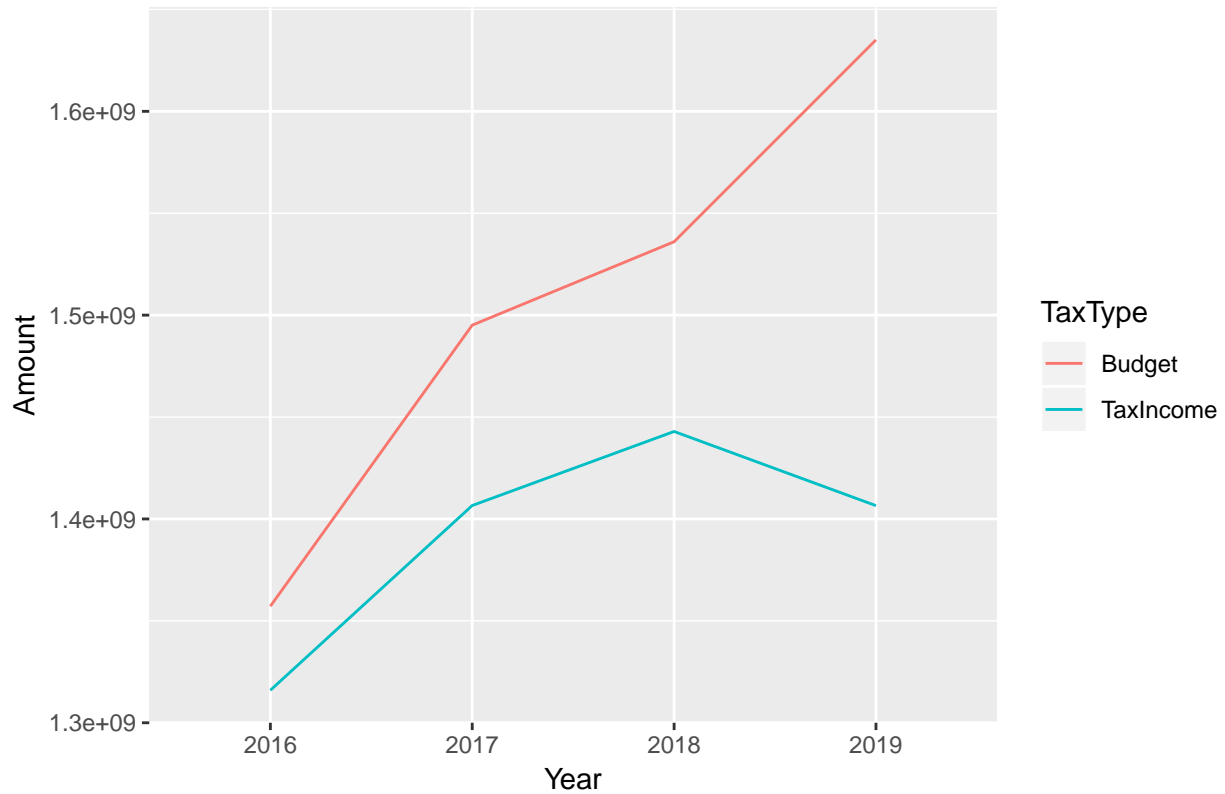
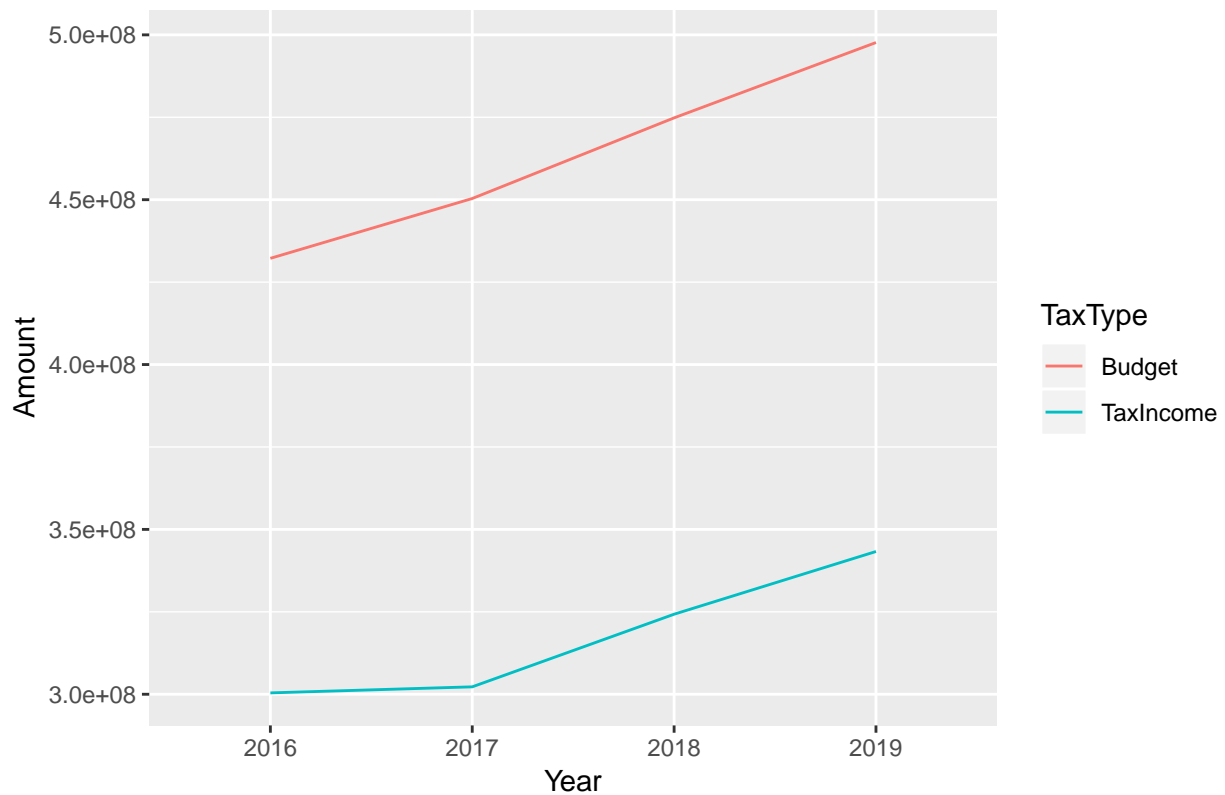


Figure 2: Budget and Tax Income v.s. Year of Burnaby



For each year, we predicted the tax income of three class codes (01, 05, 06) in each municipality by multiplying access total and mill rate. Then we plotted the predicted tax income vs municipal budget into line-chart to visualize their relationship. We chose Vancouver and Burnaby because they were two largest municipalities, and the plots showed that the predicted tax amount was positively related to municipal budget.

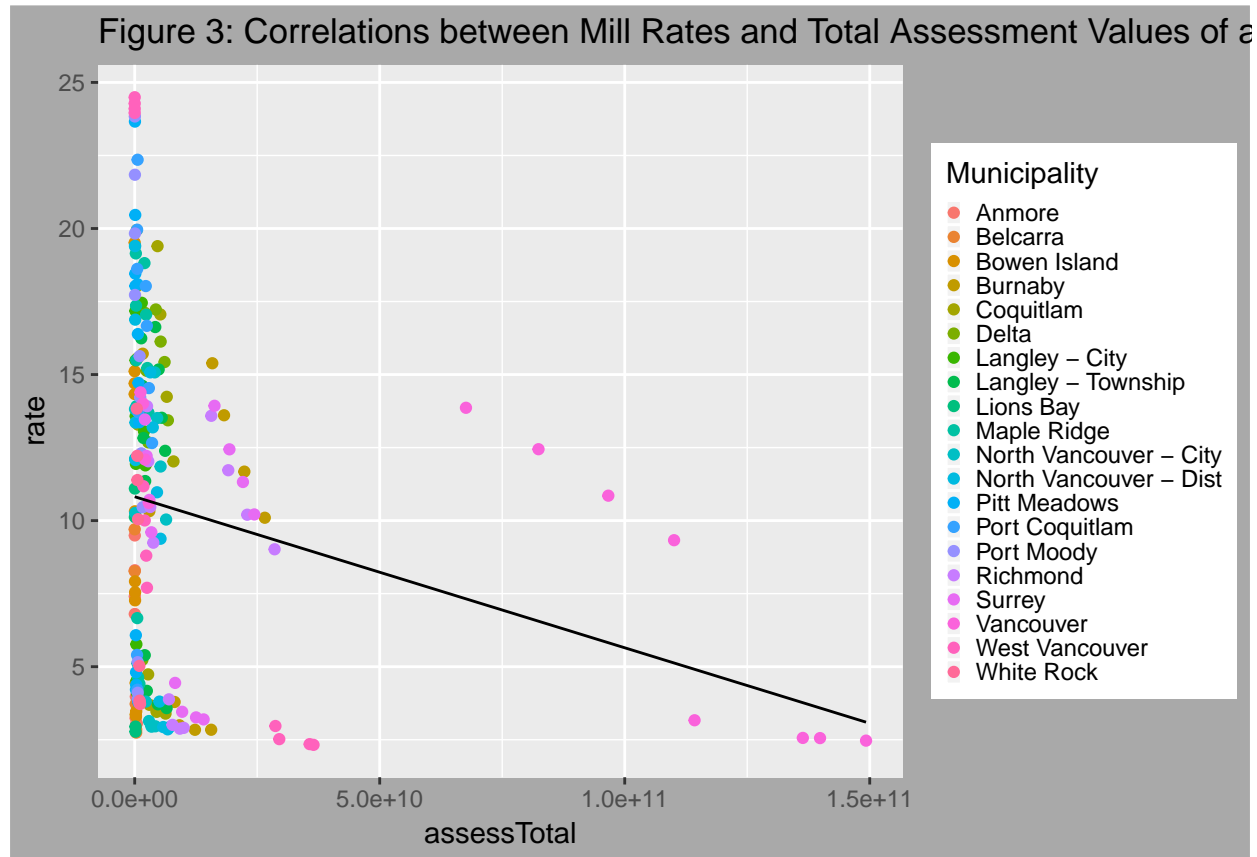


Figure 4: Correlations between Mill rates and Budgets of all municipalities

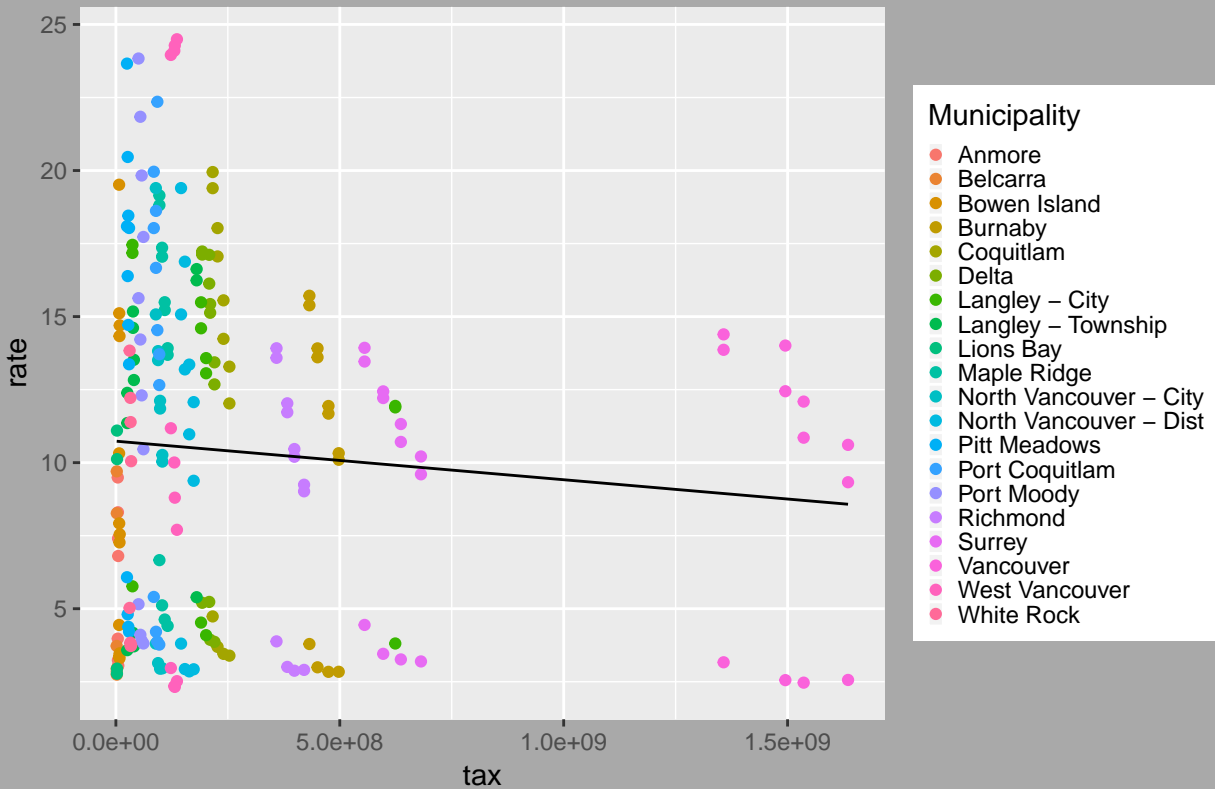


Figure 5: Correlations between Mill Rates and Total Assessment Values of L

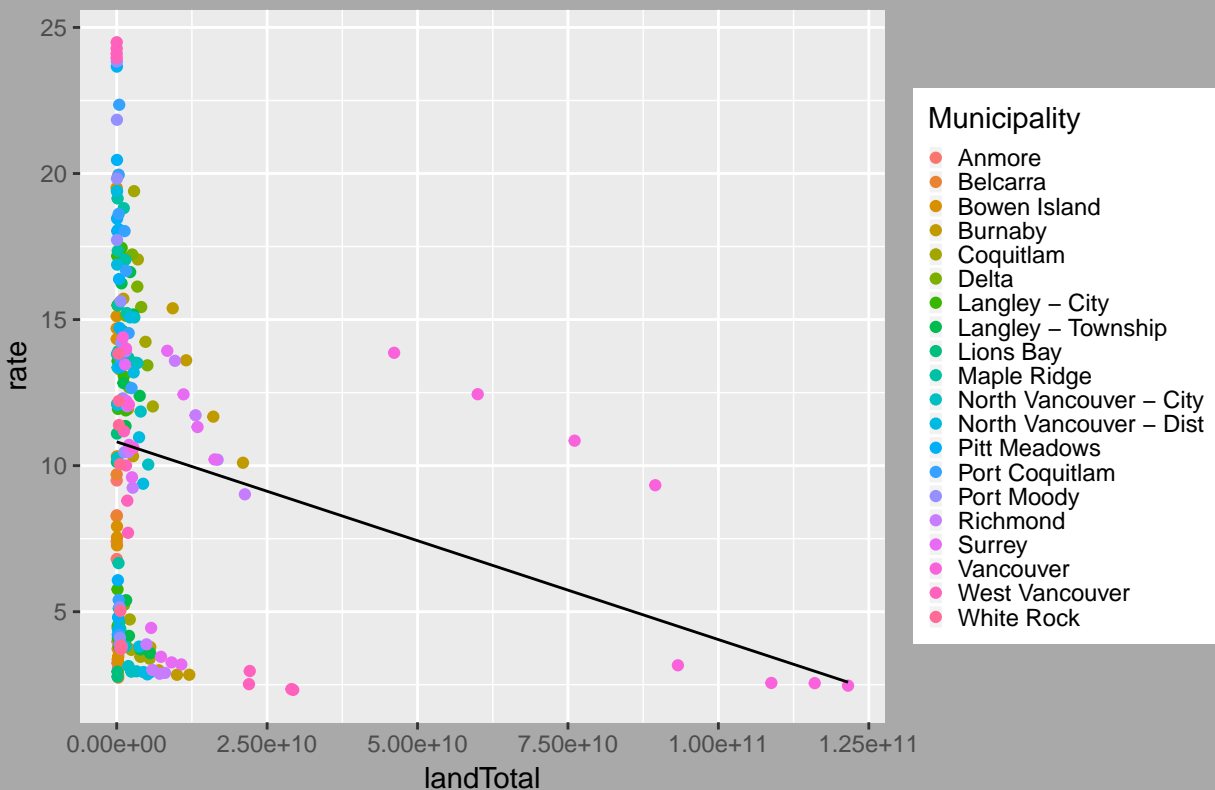


Figure 6: Correlations between Mill Rates and Total Assessment Values of I

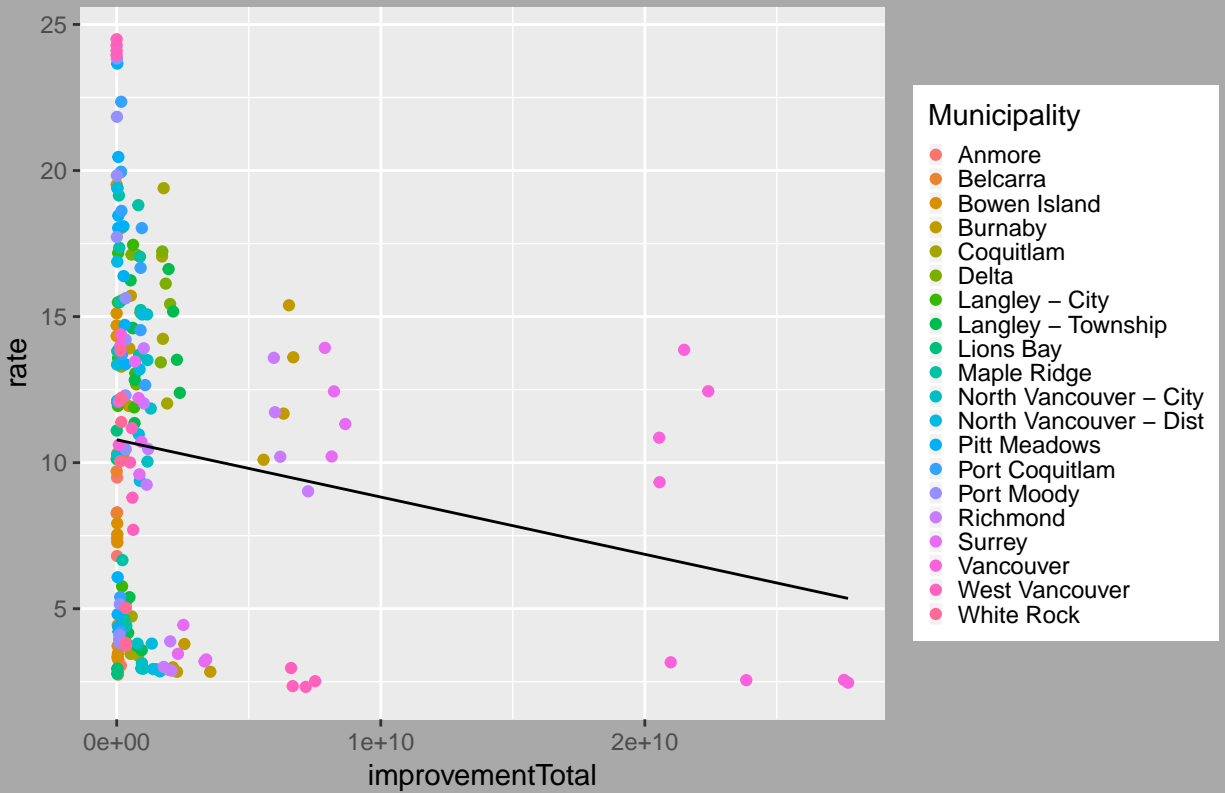
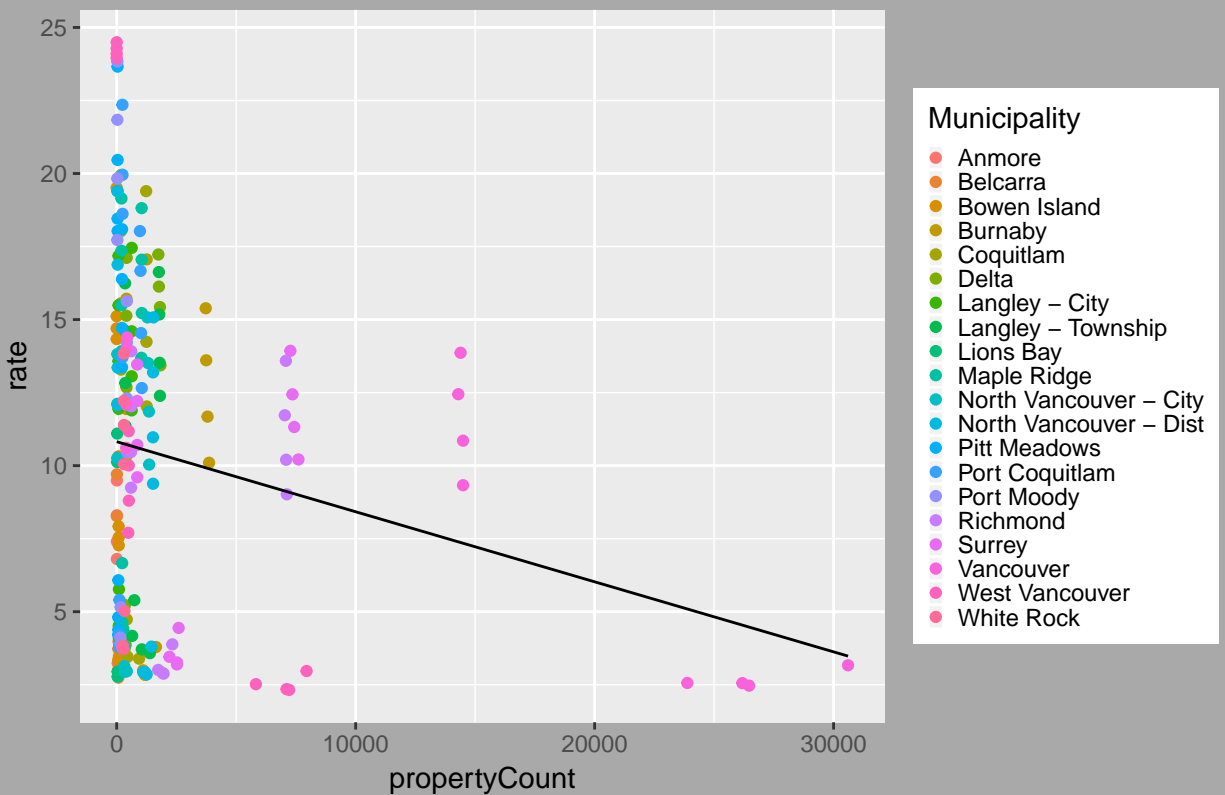


Figure 7: Correlations between Mill Rates and Total Number of Properties of



These 5 scatter plots visualized the correlation between mill rate and 5 numerical features (assessment total, municipal budget, land assessment total, improvement assessment total, and the number of properties). In every scatter plot, Vancouver (the dark pink points) are significantly different from other points and should be regarded as an outlier/special case. There are some other municipalities (including Richmond, Surrey and Burnaby) that are also distinct in the scatter plots. We believe these municipalities also have impacts on the correlation between mill rate and its features.

Feature Name	Pr(>F)
TaxClassCode	<2e-16
Year	4.2e-06
Municipality	0.0768

We used three 3 ANOVA tables to analyze the correlation between mill rate and each of the 3 categorical features (TaxClassCode, Year, Municipalities). TaxClassCode and Year showed a strong correlation with mill rate while the correlation between Municipality and mill rate is not significant at 5% level.

Results: Linear Model

We fitted 5 models and compared the goodness of fit of each models.

Model	Mutiple R-Squared	Adjusted R_Squared	MSE	PMSE
OLR full	0.8874	0.8707	1.9843	2.5902
OLR reduced	0.8874	0.8721	1.9845	2.5237
Ridge	0.8868	NA	1.9896	2.5675
LASSO	0.8873	NA	1.9855	2.5280
Elastic Net	0.8625	NA	2.2366	2.5486

Here are the specifications of each indicator:

- **Multiple R squared:** A measure of Rsquared for models that have multiple predictor variables. It will increase when adding predictors to your model.
- **Adjusted Rsquared:** It controls against the increase of Multiple R squared and adds penalties for the number of predictors in the model. Therefore it shows a balance between the most parsimonious model, and the best fitting model.
- **MSE:** An estimator that measures the average of the squares of the errors: the average squared difference between the estimated values and the actual value.
- **PMSE:** It is the expected value of the squared difference between the fitted values implied by the predictive function and the values of the (unobservable) function (In our case, we randomly select 50 samples from the data frame as testing set to estimate the PMSE).

Except for Elastic Net model which have higher MSE and PMSE, there is no significant differences between these models.

Conclusions

Among all five models, the full linear regression model is the most ideal prediction model since it make accurate prediction without losing information. There are some aspects that we could improve/implemented in the future:

- **Cross Validation:** The fitted reduced linear model was able to make the most accurate prediction across all our selected linear models, 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.
- **Data Tranformation:** 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 R^2 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. However, there could be other data transformation applied to our data, and the relationship between mill rate and other features might not be linear. GAM (Genalized addictive model) could be used to fit non-linear data.
- **Outlier:** In most correlation scatter plots, Vancouver always is an outlier. Since it makes up large portion of the market, 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, since the P-value of ANOVA between tax class and mill rate was extremely small.

References

Links to source of data:

- Schedule 706 (<https://www2.gov.bc.ca/gov/content/governments/local-governments/facts-framework/statistics/statistics>)

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)

Appendix

Missing Value:

There are 1801 missing values in mill rate (TaxClassTaxRate). Since this is the predictor, 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, aggregate them into groups by region + classcode + year.
- For entries without mill rate, find 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 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 with 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:

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