Individual Report

Real Estate - Xuechun Lu

01/03/2020

Summary

A real estate analyst in Vancouver was interested in predicting future property tax for his company by predicting future mill rates. His expectation was to obtain a good prediction model to predict mill rates in each municipality in Metro Vancouver given the past data. In this study, data cleaning was done to prepare for the data analysis and raw data was transformed into summary statistics for each municipality. Exploratory data analysis (EDA) shows that mill rate is strongly correlated with year, municipality, tax class, assessment total, improvement, land and property count, and weakly correlated with municipal budget. Linear regression models (ordinary linear regression, Lasso, Ridge and Elastic Net regression) were built to explore the relationship between mill rate and potential features. Among all models, ordinary linear regression with reduced features gives the slightly better goodness of fit and prediction power.

Introduction

The goal of this study is to predict the 2019 mill rate for the following 21 municipalities in Metro Vancouver:

- 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

In this study, EDA was first done to visualize the trend of mill rats, and relationships between mill rate and all potential features. Based on results of EDA, we built ordinary linear regression, Lasso, Ridge and Elastic Net models and evaluated their goodness of fit (how well data fit models) using multiple R squared and adjusted R squared respectively. To further test model performance, prediction power (how well models can predict future values) for each model was examined by using mean squared prediction error (MSPE).

Data Description

The real estate analyst provided past property assessment data with 2.4 million data entries from 2016 to 2019. We filtered information about commercial and residential properties (tax class 1, 5 and 6) in all 21 municipalities in Metro Vancouver.

Variables relevant to property tax prediction are selected as below:

Variable name	Interpretation	Type
rate	Mill Rates (2016-2019)	Quantitative
assessTotal	Assessment (2016-2019)	Quantitative
landTotal	Land Total (2016 - 2019)	Quantitative
ImprovementTotal	Improvement Total (2016 - 2019)	Quantitative
propertyCount	Number of Properties (2016-2019)	Quantitative
tax	Municpal Budget	Quantitatiive
TaxClassCode	Tax Class Code (01, 05, 06)	Categorical
AddressAssessorMunicipalityDesc	Area Code (Municipality)	Categorical
Year	Year (2016, 2017, 2018, 2019)	Categorical

External Data:

The analyst client also proposed that municipal budget could be related to mill rate, however, this data was not available neither in the data set he provided or online. Therefore, we collected external data from the Government of British Columbia to approximate the municipal budget: Municipal Budget of Cities in Metro Vancouver (Taxes Imposed & Collected, Schedule 706)

The data listed were all cleaned to reduce their dimensions before data analysis and were all aggregated to "assessment_aggregate.csv".

Method

1. Exploratory Data Analysis (EDA)

Before model construction and prediction of mill rates, EDA needs to be done to visualize the main characteristics of our dataset. In this study, EDA mainly focuses on exploring the relationships between mill rate and potential features. We first plotted correlations between mill rate and quantitative features (assessTotal, improvementTotal, propertyCount and landTotal). As for categorical features (TaxClassCode, Year and AddressAssessorMunicipalityDesc), instead of plotting correlations for all combination of them (3*4*21), ANOVA analysis was used to test if the correlation between mill rate and each categorical feature is significant.

2. Measure of goodness of fit and prediction power

In this study, we first built various models. The evaluation of model performances is based on goodness of fit (how well the model explains the data) and prediction power (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. Multiple R squared and adjusted R squared are the most well known measures of goodness of fit and higher multiple R squared and adjusted R squared generally indicate better goodness of fit. Moreover, for ordinary linear regression (OLR), reduction of number of features decreases multiple R squared. Therefore, when comparing OLR with different number of features, adjusted R squared will be looked at, instead of R squared.

Prediction power measures how well models can predict the future values. We used mean squared prediction error (MSPE) to examine prediction performances across all the models. MSPE is the mean squared prediction error calculated from the testing set. To obtain it, we divided the data set into training data and testing data, which would be used as model construction and prediction respectively. Generally, smaller MSPE indicates better prediction power.

3. Ordinary Linear Model

The features we are interested in are AddressAssessorMunicipalityDesc (municipality), Year, TaxClassCode, assessTotal, landTotal, improvementTotal, propertyCount, predTaxIncome and tax (municipal budget).

We first built the ordinary linear model using all features available (OLR full) and computed its R squared, adjusted R squared and MSPE on the whole data set as measures of goodness of fit. Next, based on its R output, we selected the significant variables and re-built a OLR model using only these variables (OLR reduced), and we computed goodness of fit by repeating the previous step.

4. Ridge, Lasso and Elastic Net

Except for ordinary linear regression, we are also interested in the performance of advanced linear regressions like Ridge, Lasso and Elastic Net as they are also in the linear regression family, but have a different objection function to optimize. Among the three models, weights are assigned to features, and we are interested in minimizing the sum of squares of errors plus a penalty term. Ridge regression takes penalty in sum of absolute values of weights, whereas Lasso takes penalty in sum of squared absolute values of weights. Elastic net is a combination of Ridge and Lasso.

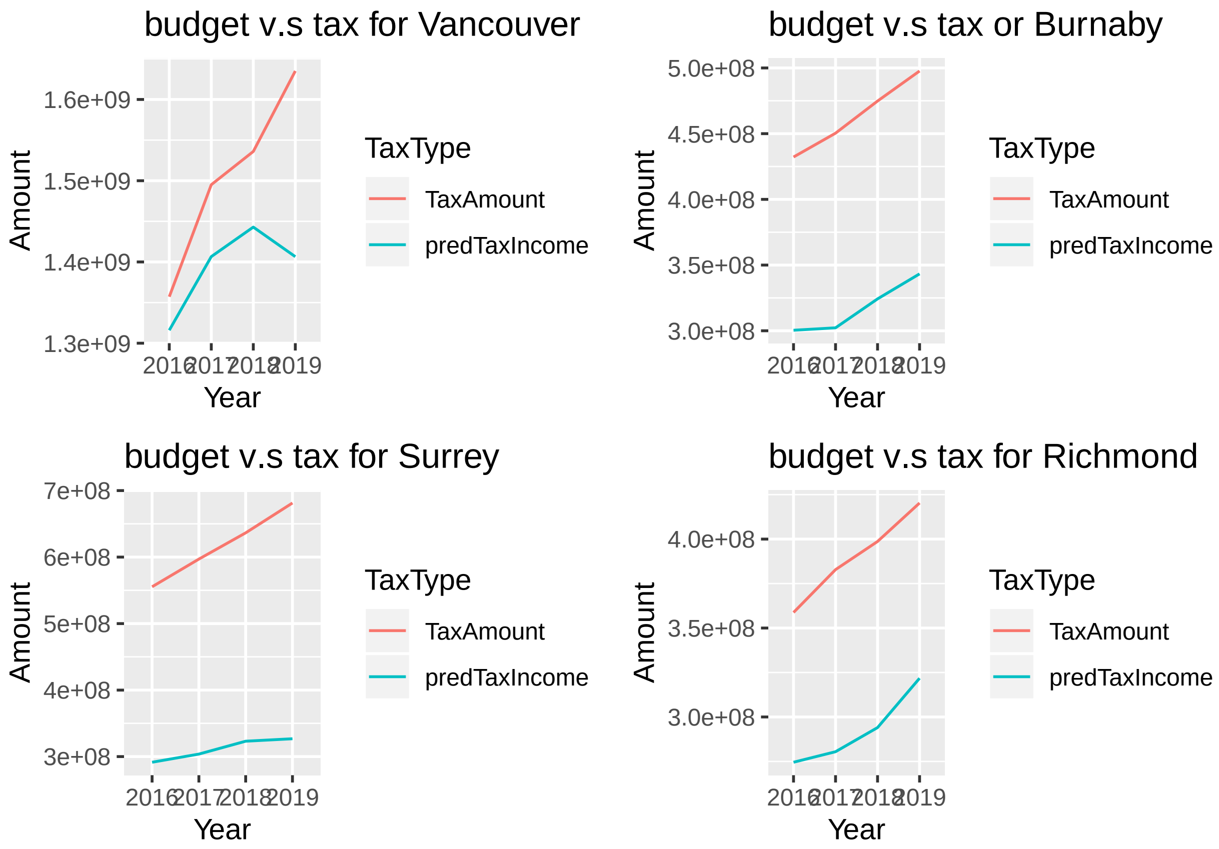
The reason for considering advanced linear regressions is that assigning weights to features might improve the goodness of fit and prediction power as each feature might influence the mill rate to a different extent.

Similar to the model construction for OLR, we first built the three models using all data and evaluated their goodness of fit. Then we split the data into training and testing sets and computed MSPE to examine their prediction power.

Results

1. EDA

```
# get predicted tax income for all tax classes and each municipality
assessment_aggregate[,10] <- assessment_aggregate$assessTotal*assessment_aggregate$rate/1000
names(assessment_aggregate)[10] <- paste("predTaxIncome")

dat_pred_income <- assessment_aggregate %>% select(Year,AddressAssessorMunicipalityDesc,TaxClassCode,pr
dat_pred_income <- aggregate(dat_pred_income$predTaxIncome,by=list(Year=dat_pred_income$Year,Municipali
names(dat_pred_income)[3] <- paste("predTaxIncome")

tax_refactored <- tax_modified
names(tax_refactored)[3] <- paste("2016")
names(tax_refactored)[4] <- paste("2017")
names(tax_refactored)[5] <- paste("2018")
names(tax_refactored)[6] <- paste("2019")
dat_true_income <- gather(tax_refactored, Year, TaxAmount, "2016":"2019", factor_key=TRUE)
dat_true_income <- dat_true_income[,-1]

income_budget_compare <- merge(dat_true_income,dat_pred_income,by=c("Municipalities","Year"))
income_budget_compare <- gather(income_budget_compare, TaxType, Amount, TaxAmount:predTaxIncome, factor
compare_full_plot <- income_budget_compare %>% filter(Municipalities!="Vancouver") %>% ggplot(aes(x=Yea

compare_plot1 <- income_budget_compare %>% filter(Municipalities=="Vancouver") %>% ggplot(aes(x=Year,y=A
compare_plot2 <- income_budget_compare %>% filter(Municipalities=="Surrey") %>% ggplot(aes(x=Year,y=Amou
compare_plot3 <- income_budget_compare %>% filter(Municipalities=="Burnaby") %>% ggplot(aes(x=Year,y=Am
compare_plot4 <- income_budget_compare %>% filter(Municipalities=="Richmond") %>% ggplot(aes(x=Year,y=A
multiplot(compare_plot1,compare_plot2,compare_plot3,compare_plot4,cols=2)
```

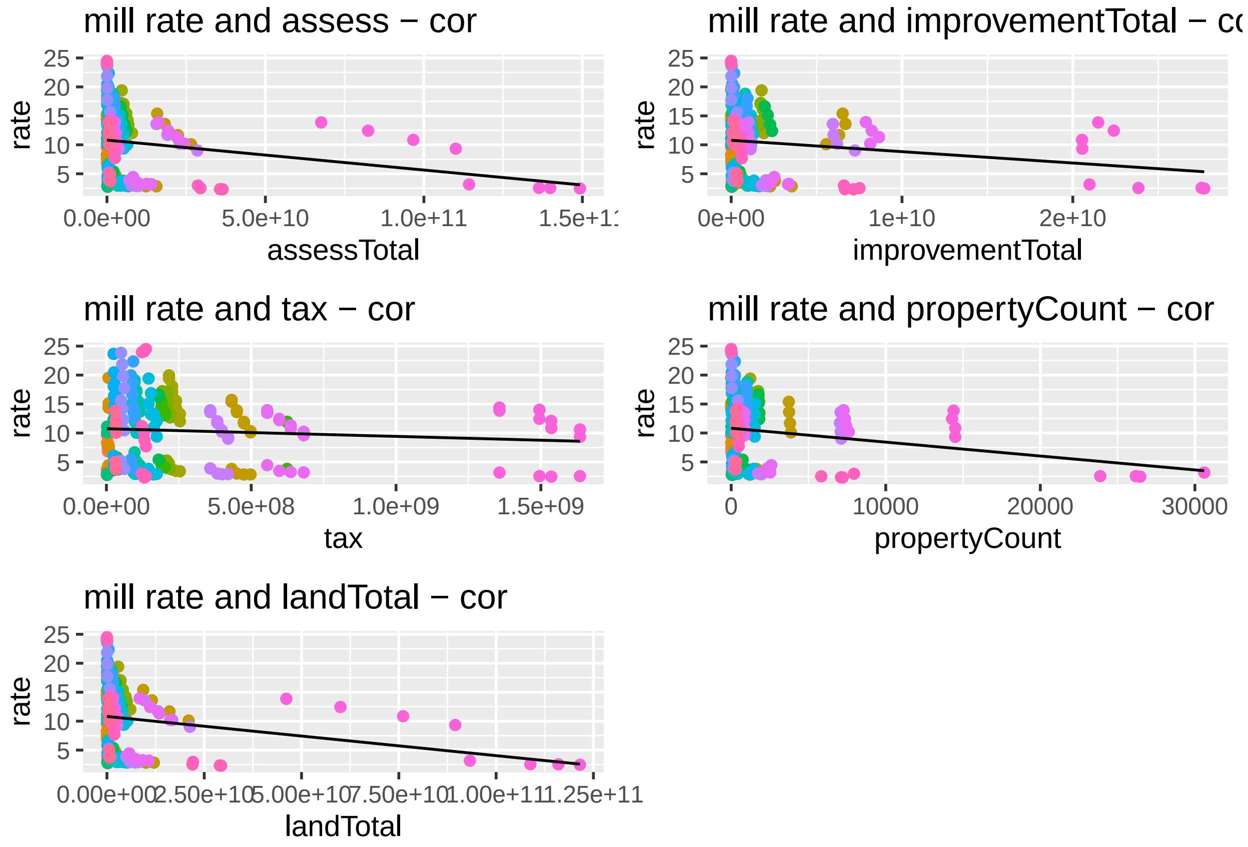


Plot 1. budget and tax to pay from 2016-2019 in 4 major municipalities

We first plot municipal budget (taxAmount) and predicted property tax to pay (predTaxIncome) against time from 2016 to 2019 in 4 major municipalities (Vancouver, Burnaby, Surrey and Richmond). In the plot above, the predicted tax to pay is less than the municipal budget. Also, budget and predicted tax generally have an increasing trend and move together in the consecutive four years, which indicates a potential correlation between the two variables.

```
# aggregate tax and mill rate & assessment
colnames(tax_modified) <- c("X","Municipalities","2016","2017","2018","2019")
dat_tax_long <- gather(tax_modified, Year, TaxAmount, "2016":"2019", factor_key=TRUE)

tax_modified1 <- gather(tax_modified, Year, tax, '2016':'2019', factor_key = TRUE)%>%
  rename(AddressAssessorMunicipalityDesc = Municipalities)

assessment_aggregate <- merge(assessment_aggregate, tax_modified1, by = c("AddressAssessorMunicipalityDe
assessment_aggregate <- assessment_aggregate[,-c(3,11)]
names(assessment_aggregate)[10] <- paste("tax")
```

```
## Correlation for continuous variables
assess_class <- assessment_aggregate %>% ggplot(aes(x=assessTotal,y=rate,group=AddressAssessorMunicipal

tax_class <- assessment_aggregate %>% ggplot(aes(x=tax,y=rate,group=AddressAssessorMunicipalityDesc,col

land_class<-assessment_aggregate %>% ggplot(aes(x=landTotal,y=rate,group=AddressAssessorMunicipalityDes

improvementTotal_class<-assessment_aggregate %>% ggplot(aes(x=improvementTotal,y=rate,group=AddressAsse


propertyCount_class<-assessment_aggregate %>% ggplot(aes(x=propertyCount,y=rate,group=AddressAssessorMu

multiplot(assess_class,
tax_class,
land_class,improvementTotal_class,
          propertyCount_class, cols=2)
```



Plot 2. scatterplot with regression line for mill rate against assessTotal, tax, landTotal, improvementTotal and propertyCount.

Plot 2 shows the correlation between mill rate and five quantitative variables that likely affect the mill rate. Mill rate is negatively correlated with assessTotal, improvementTotal, propertyCount and landTotal considering all the 21 municipalities in Metro Vancouver, as the black linear regression lines all have a negative slope. In contrast, mill rate and budget is weakly correlated since the linear regression line is almost flat.

Table 2: ANOVA analysis of mill rate and three categorical variables

variable	df	sum sq	mean sq	F value	Pr(>F)
factor(TaxClassCode)	2	5796	2897.9	340.6	<2e-16 ***
factor(Year)	3	405	135.13	4.006	0.00841 **
factor(AddressAssessorMunicipalityDesc)	19	979	51.53	1.535	0.0768 .

```
## [1] "Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1"
```

Table 2 lists the result of ANOVA analysis of mill rate and three categorical variables - TaxClassCode, Year and Municipality (AddressAssessorMunicipalityDesc). A variable with p-value (Pr(>F)) less than 5% is considered statistically significant at 5% significance level. From the table, TaxClassCode and Year are statistically significantly correlated with mill rate at 5% level.

2. Goodness of Fit

```
# full linear model
linear_full<-lm(rate~factor(AddressAssessorMunicipalityDesc)+factor(Year)+factor(TaxClassCode)+assessTo
summary(linear_full)
linear_full_fit<-augment(linear_full)
sqrt(sum((linear_full_fit$.resid)^2)/nrow(assessment_aggregate))
# Multiple R-squared:  0.8874,  Adjusted R-squared:  0.8707
# MSPE = 1.984285
row1<-c("OLR full",0.8874, 0.8707,round(1.984285,4))

# reduced model
reduced<-lm(rate~factor(Year)+factor(TaxClassCode)+factor(AddressAssessorMunicipalityDesc)+assessTotal+
summary(reduced)
reduced_fit<-augment(reduced)
# Multiple R-squared:  0.8874,  Adjusted R-squared:  0.8721
sqrt(sum((reduced_fit$.resid)^2)/nrow(assessment_aggregate))
row2<-c("OLR reduced",0.8874,0.8721,round(1.984541,4))
```

Table 3: List of significant variables in OLR full

Variable	Pr(>|t|)	Significance level
Intercept	0.000111	***
Coquitlam	0.001594	**
Delta	0.002579	**
Langley - City	0.020365	*
Langley - Township	0.023030	*
Maple Ridge	4.18e-05	***
Pitt Meadows	2.41e-07	***
Coquitlam	1.60e-05	***
Port Moody	1.16e-05	***
West Vancouver	0.001003	**
White Rock	0.034041	*
Year 2017	0.000655	***
Year 2018	1.30e-08	***
Year 2019	4.73e-13	***
TaxClassCode 5	< 2e-16	***
TaxClassCode 6	< 2e-16	***

Variable	Pr(>|t|)	Significance level
assessTotal	0.013752	*
landTotal	0.013863	*

```
## [1] "Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1"
```

```
# Ridge
dummy_year<-dummy(assessment_aggregate$Year)
dummy_municipal<-dummy(assessment_aggregate$AddressAssessorMunicipalityDesc)
dummy_taxclass<-dummy(assessment_aggregate$TaxClassCode)
# build x matrix
x<-cbind(dummy_municipal,dummy_year,dummy_taxclass,assessment_aggregate$assessTotal,assessment_aggregate
y<-assessment_aggregate$rate
lambdas <- 10^seq(2, -3, by = -.1)

lambdas <- 10^seq(2, -3, by = -.1)
ridge_reg = glmnet(x, y, nlambda = 25, alpha = 0, family = 'gaussian', lambda = lambdas)
set.seed(450)
cv_ridge <- cv.glmnet(x, y, alpha = 0, lambda = lambdas, nfolds=10)
optimal_lambda <- cv_ridge$lambda.min
optimal_lambda
predictions_train <- predict(ridge_reg, s = optimal_lambda, newx = x)

# Compute R^2 from true and predicted values
eval_results <- function(true, predicted) {
  SSE <- sum((predicted - true)^2)
  SST <- sum((true - mean(true))^2)
  R_square <- 1 - SSE / SST
  MSPE = sqrt(SSE/nrow(predicted))
# Model performance metrics
data.frame(
  MSPE = MSPE,
  Rsquare = R_square
)

}

predictions_train <- predict(ridge_reg, s = optimal_lambda, newx = x)
eval_results(y, predictions_train)

row3<-c("Ridge",round(0.8868242,4),"N/A",round(1.989583,4))


# Lasso
set.seed(450)
lasso_reg <- cv.glmnet(x, y, alpha = 1, lambda = lambdas, standardize = TRUE, nfolds = 10)
lambda_best <- lasso_reg$lambda.min;lambda_best
lasso_model <- glmnet(x, y, alpha = 1, lambda = lambda_best, standardize = TRUE)
predictions_train <- predict(lasso_model, s = lambda_best, newx = x)
eval_results(y, predictions_train)

row4<-c("Lasso",round(0.8872924,4)  ,"N/A",round(1.985463,4))

# Elastic Net
```

```r
#tibble::as_tibble(assessment_aggregate)
cv_10 = trainControl(method = "cv", number = 10)
elastic_net = train(
  rate~factor(AddressAssessorMunicipalityDesc)+factor(Year)+factor(TaxClassCode)+assessTotal+landTotal+
  method = "glmnet",
  trControl = cv_10
)
elastic_net
row5<-c("Elastic Net",round(0.8625398,4),"N/A",round(2.236598,4))
```

Table 4: Goodness of fit for all models

Model	Multiple R-squared	Adjusted R-squared	MSPE
OLR full	0.8874	0.8707	1.9843
OLR reduced	0.8874	0.8721	1.9845
Ridge	0.8868	N/A	1.9896
Lasso	0.8873	N/A	1.9855
Elastic Net	0.8625	N/A	2.2366

We performed ordinary linear regression (OLR) with all variables first and computed multiple R-squared, adjust R-squared and mean square prediction error (MSPE). Since the summary statistics in Table 3 show that only Municipality, Year, TaxClassCode, assessTotal and landTotal are significant, we therefore constructed a reduced OLR model with only the significant variables.

Table 5 shows the comparison of the full model (OLR full) and the reduced model (OLR reduced). Multiple R-squared doesn't seem to improve, but adjusted R-squared improves by a small amount. This indicates that with fewer variables, the goodness of fit improves slightly. In addition, with the three advanced linear models (Ridge, Lasso and Elastic Net), the family of OLR have slightly better fit and OLR reduced has the best fit as its adjusted R-squared is slighter higher than OLR full.

3. Prediction Power

```r
set.seed(450)
train_ind<-sample(218,218-50)
train<-assessment_aggregate[train_ind,]
test<-assessment_aggregate[-train_ind,]
```

```r
# full linear model
newx<-test[,-c(8,9)]
y<-test[,c(8)]
linear_1<-lm(rate~factor(AddressAssessorMunicipalityDesc)+factor(Year)+factor(TaxClassCode)+assessTotal
resid<-predict(linear_1,newdata = newx) - y
row1<-c("OLR full", round(sqrt(sum(resid^2)/nrow(test)),4))

# reduced model
linear_2<-lm(rate~factor(Year)+factor(TaxClassCode)+factor(AddressAssessorMunicipalityDesc)+assessTotal
resid<-predict(linear_2,newdata = newx) - y
row2<-c("OLR reduced",round(sqrt(sum(resid^2)/nrow(test)),4))
```

```r
# lasso
# create the whole matrix
y<-as.matrix(assessment_aggregate$rate)
dim(x) # 165  29
dim(y)
# creat x_train matrix and y_train
x_train<-x[train_ind,]
y_train<-y[train_ind,]
# create x_test matrix
x_test<-x[-train_ind,]
y_test<-y[-train_ind,]

# Setting alpha = 1 implements lasso regression
set.seed(450)
lasso_reg <- cv.glmnet(x_train, y_train, alpha = 1, lambda = lambdas, standardize = TRUE, nfolds = 10)
lambda_best <- lasso_reg$lambda.min;lambda_best
lasso_model <- glmnet(x_train, y_train, alpha = 1, lambda = lambda_best, standardize = TRUE)
predictions_test <- predict(lasso_model, s = lambda_best, newx = x_test)
eval_results(y_test, predictions_test)
row4<-c("Lasso",round(2.528047,4))


# ridge
ridge_reg = glmnet(x_train, y_train, nlambda = 25, alpha = 0, family = 'gaussian', lambda = lambdas)
set.seed(450)
cv_ridge <- cv.glmnet(x_train, y_train, alpha = 0, lambda = lambdas, nfolds=10)
optimal_lambda <- cv_ridge$lambda.min
optimal_lambda
predictions_test <- predict(ridge_reg, s = optimal_lambda, newx = x_test)
eval_results(y_test, predictions_test)
row3<-c("Ridge",round(2.567499,4))

# elastic net
tibble::as_tibble(assessment_aggregate[train_ind,])
cv_10 = trainControl(method = "cv", number = 10)
elastic_net = train(
 rate~factor(AddressAssessorMunicipalityDesc)+factor(Year)+factor(TaxClassCode)+assessTotal+landTotal+in
 data = assessment_aggregate[train_ind,],
  method = "glmnet",
  trControl = cv_10
)
elastic_net

# RMSE was used to select the optimal model using the smallest value.
# The final values used for the model were alpha = 1 and lambda = 0.06549203.

elastic_reg = glmnet(x_train, y_train, nlambda = 25, alpha = 1, family = 'gaussian', lambda =  0.0654920
predictions_test <- predict(elastic_reg, newx = x_test)
eval_results(y_test, predictions_test)
row5<-c("Elastic Net",round(2.54861,4))
```

Table 5: Prediction power for all models

Model	MSPE
OLR full	2.5902
OLR reduced	2.5237
Ridge	2.5675

Model	MSPE
Lasso	2.528
Elastic Net	2.5486

Table 5 shows that among the five models, OLR reduced has the smallest MSPE and thus the best prediction power. However, the difference between OLR reduced and other models is not big.

Conclusion

The OLR shows significant variables related to mill rates are Year, Tax Class, Municipality, Assessment and Land Total

OLR reduced has slightly better goodness of fit of the model and prediction power among all models. However, this good fit is likely to be a result of over-fitness since for each municipality and tax class, we only have 4 consecutive data points from 2016 to 2019.

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). 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 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:

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