**Housing Price Analysis and Prediction**

**ALY 6040 Data Mining Applications**

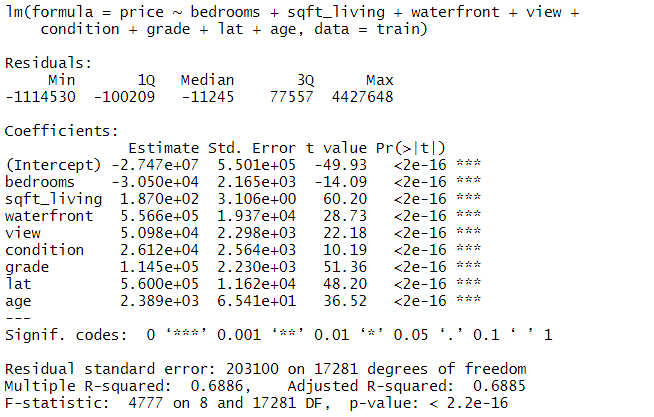
**Week 5 Assignment - Optimization Techniques**

**Northeastern University**

**Spring 2019**

**Review**

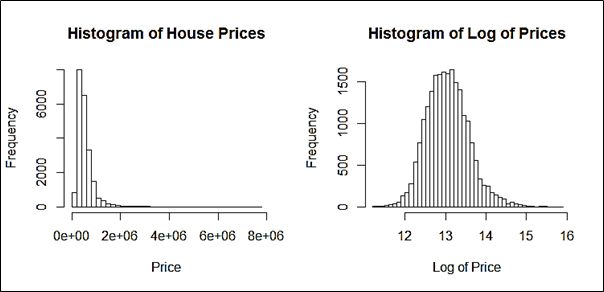
We chose using linear regression to predict the future housing price in King Country timely and accurately as time goes on. We set ‘price’ as the dependent variable and others as independent variables. Most of them are significantly correlated with each other. In total, we excluded a few variables because they have higher P-values than 0.05 such as ‘bathroom’ and ‘longitude’. The variables that are maintained in the table below have P-values lower than 0.05, which means they are significant.

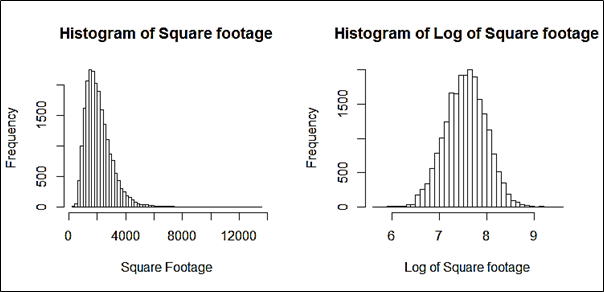
****

In addition, we also observed R-squared, MAE, and RMSE for linear regression. Since the MAE and RMSE are quite high, we need to optimize the model by using some optimization techniques. Then, we come up with some business learnings which are important for the buyers, government, and real-east investors to pay attention if they want to predict the housing price.

**Variables Transformations**

The histograms of variables in the dataset show that 'price' (price of houses) and 'sqft\_living' (square footage of houses) follow a log normal distribution. The variables were transformed using the logarithmic function, and the resultant variables follow an approximated normal distribution.



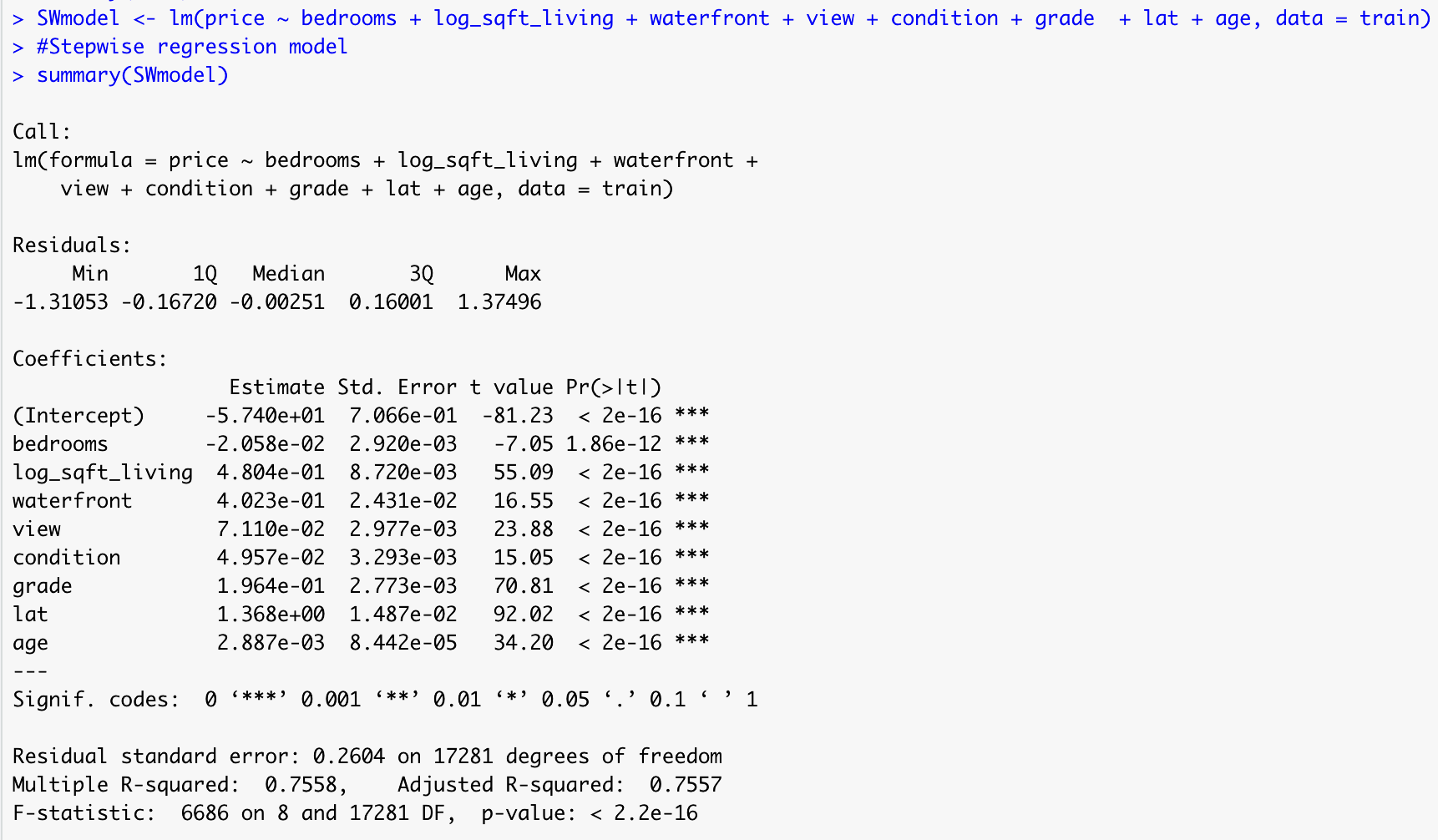


By transforming the two variables using logarithmic function, the efficiency of the baseline linear model increases substantially. The R2 of the baseline linear regression model was observed to be 0.69. However, after performing the transformations, the R2 of the linear regression model improved to 0.76.

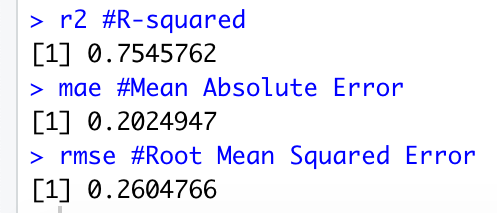
**Optimization Techniques**

**Stepwise**

Stepwise regression is a combination of using the forward technique (adding new variables) and backward elimination (removing any variables) to improve the model to a statistically significant extent.



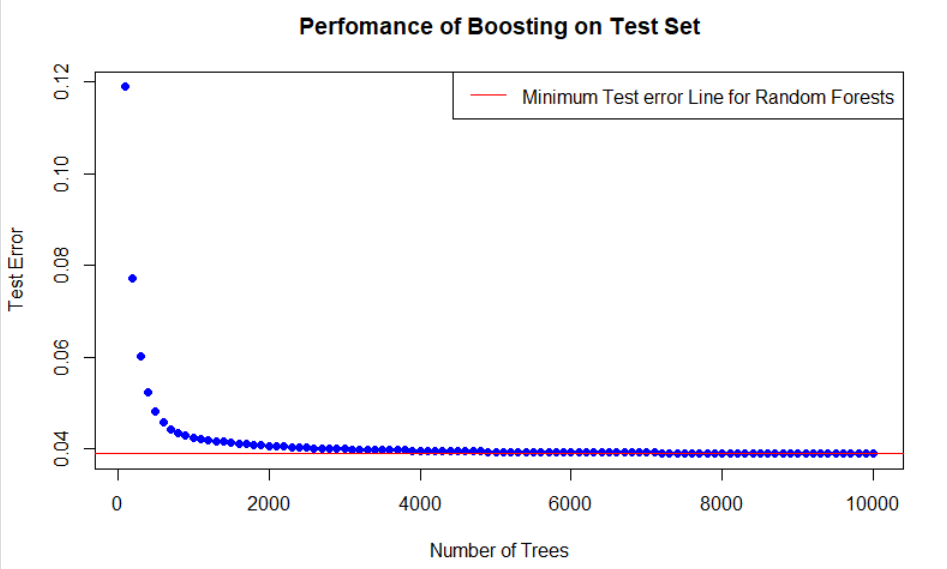
In this project, we generate Stepwise model using 8 variables: bedrooms, log\_sqft\_living, waterfront, view, condition, grade, lat, and age. As we can see, all 8 variables have a significant effect on the price of houses because of P-value of each variable smaller than 0.05.



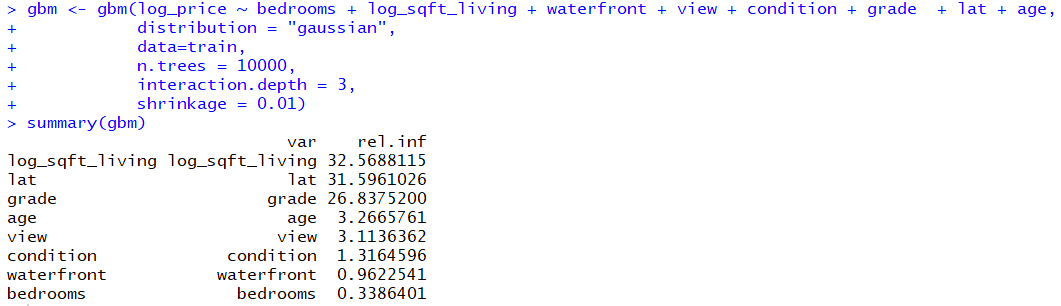
The results show that R-squared value is 0.755 with the Root Mean Squared Error = 0.260 and the Mean Absolute Error is 0.202.

**Gradient Boosting**

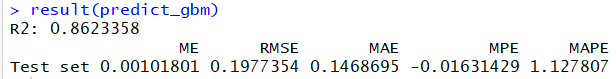
Gradient Boosting Classifier (GBC) is a helpful algorithm to create a prediction model. In this case, we try to use GBC to build the model based on the same variables with the original model. In this case, I try to determine the test error of the model with the number of trees varying from 0 to 10000.



The graph shows that the higher number of trees, the smaller test errors. Thus, we choose 10000 for the number of trees and generate a gradient boosting model with the training dataset by using the interaction depth = 3 and the shrinkage = 0.01.



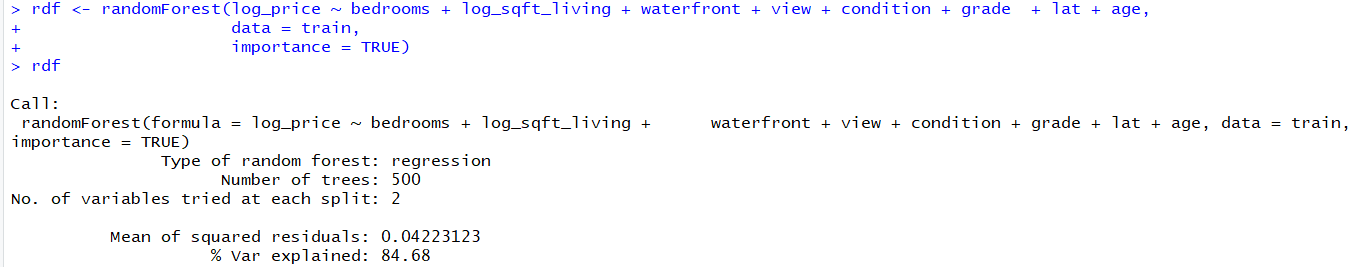
The results collected from this model show that square feet living is the variable having the highest relative influence out of others, while the number of bedrooms is not related to the house prices in this dataset with just only 0.3 points.

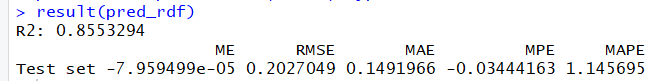


It is clear to see that Gradient boosting model has the R-squared value = 0.86 with the RMSE = 0.2 and MAE = 0.14.

**Random Forest**

Random forest is the model which is suitable for the dataset having large numbers of features. It will create some decision trees from the dataset and evaluate the best trees to use for the upcoming model. In this case, we continue using the same independent variables with the linear regression model to determine the differences between the two models.

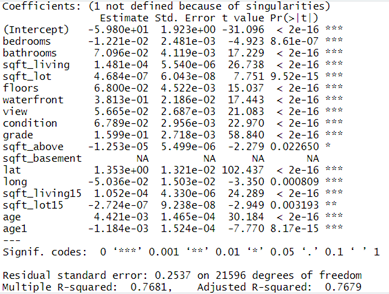


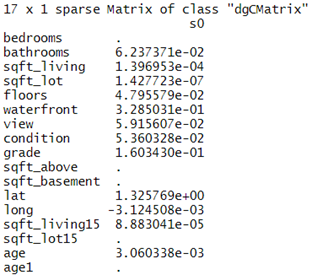
****

The results show that the R-squared value of the random forest model is 0.85 with the RMSE = 0.2 and the MAE = 0.15.

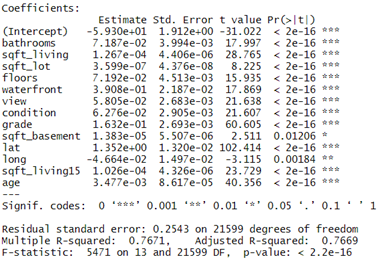
**Lasso Regression**

We tried to use the method of Regularization for optimization to reduce the error and improve accuracy. To use the Lasso regression, called L1, we removed the meaningless variable and used the remaining 18 variables. We use the “glmnet” function to use the lasso regression.



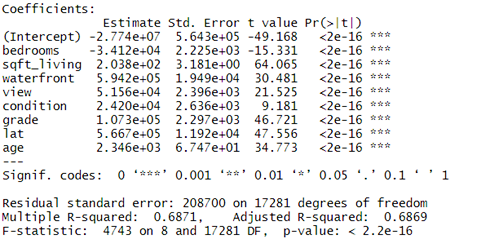
This is the result of a linear regression analysis using the original 18 variables before using the Lasso regression. The age1 variable, which means the time of sale after the house was renovated, is discovered as only one meaningless variable due to its high p-value. The R-squared of this analytical model is 0.7681 and the Adjusted R-squared is 0.7679. We conducted the Lasso regression to see if the Lasso Regression could help improve the accuracy of the model. We used the “cv.glmnet” function and set the lambda value to 500.

With the seed set to 123 and Lasso done, the variables bedrooms, sqft\_above, sqft\_basement, sqft\_lot15, and age1 variables were cleared. We examined whether the accuracy of linear regression could be improved by using 12 independent variables without 5 variables removed above.

We used the Lasso regression to clear the five variables and attempted to rebuild the model again, but the results were almost identical. The p-value of each variable did not change and the r value which means the explanatory power of the model decreased slightly. Therefore, we decided that Lasso Regression did not help improve the accuracy of our model.

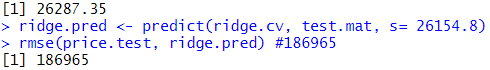
**Ridge Regression**

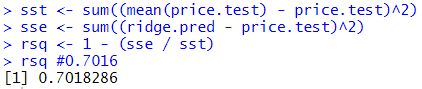
We have tried Ridge Regression because Lasso regression was not helpful in improving the accuracy of the model. Since we did not see any improvement in the model when we used 8 variables for Lasso regression, we will only use 8 independent variables for ridge regression.





The results of the regression analysis using 8 variables showed that all of the variables were significant and that R-squared was 0.6871 and adjusted R-squared was 0.6869. We went through the tests to see what improvements the Ridge Regression could do. We found the best lambda value before running the ridge regression, which is 24922.27. We run the test using the best lambda value and the rmse value is 186965, which is decreased by 1068 before the ridge regression was performed.

Also, the R-squared obtained by performing the test after using the ridge regression increased by about 0.2 points as 0.7018.



Therefore, we can say that the ridge regression helps to reduce the error of the model slightly and increases the accuracy by about 0.2.

**Comparison and Findings for Business Use**

The R2 values of Stepwise, Gradient Boosting and Random Forest models are 0.77, 0.862, and 0.855 respectively. This indicates that Gradient Boosting and Random Forest perform substantially better than the Stepwise model. However, Gradient Boosting performs relatively better than Random Forest. In addition, Random Forest requires more time and space to build a model as compared to Gradient Boosting. Hence, the best model out of the three is Gradient Boosting. The R2 for the baseline model is 0.69, and after optimization, the R2 is observed to be 0.86 (using transformation and Gradient Boosting). Thus, the business should use this model for efficient predictions.