**Saratoga house prices**

**Hand build Model**

Our group has tried several models in order to minimize the RMSE, and for each model, we found 200 times trained group and test group and workout it 200 times RMSE and average them, in order to minimize the error. (If only run the model for only 1 time, there is a big possibility of random). After testing several models’ RMSE, finally, we found a satisfied model which we think it has a low RMSE, at least the RMSE of the model we used is much lower than “medium” model. The model we used is shown below:

**price=bedrooms+ newConstruction+ heating+ fireplaces+ livingArea+ age+waterfront+centralAir+rooms\*heating+landValue\*lotSize+rooms\*bathrooms.**

And in order to find the strong drivers of house prices, I think there are two points of view to find the strong drivers. Firstly, it is important to select if you drop any of the variable in your model, your RMSE will be much bigger. (Say your model is less correctly predict than before). Secondly, check the coefficient of each variable, this aim is to see which variable’s change can lead to big change in price.

As for the first point, our group select some variables which we think can lead to big increase in RMSE, the variable we test respectively are room\*bathrooms, age, bedrooms, and centralAir. And we workout colmeans of each variable. The result is shown below (every time run the coding the result will be different, because the trained group and test group are random assigned):



v1-meidium model

v2-big model

v3-standard model-the model we use

v4-v7 are the models to test which variables in v3 model are “strong drivers”

v4-standard model without room\*bathrooms variable

v5-standared model without age variable

v6-standard model without bedrooms variable

v7-standard model without centralAir Variable

From the result above we can see standard model has the lowest RMSE. (cause it runs 200 times so the colMeans result is stable to most extent) And it is much lower than the medium model.

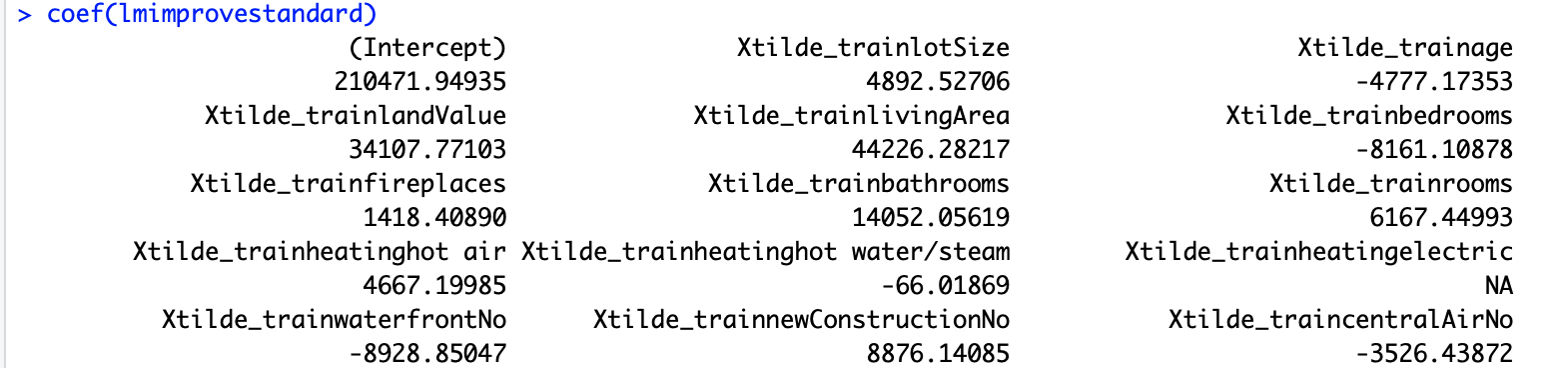
Comparing v3 with v4-v7, we can see the biggest different is v7(without centralAir variable) the second biggest difference is V6 (without bedrooms variable) So accordingly, we think **centralAir and bedrooms are strong drivers house prices according to RMSE indicator.**

As for the second point, our group will workout the coefficient of the model we select, but it is not right to directly to find out their coefficient, for example: for living area and bedrooms, adding 1 more feet of living area is different from adding 1 more bedroom. In this case, we standardized each variable. Using the formula shows below:

图片包含 物体

描述已自动生成

After standardrized, we found the coefficient of each variable:



And we definite the strong driver as the top 3 coefficient variables are strong driver: so in this case, livingarea, bathrooms and landValue are 3 strong drivers.

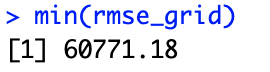
**KNN Model**

When using KNN model, we just follow the instruction, don’t include interactions of our standard model, and after standardized, we plot RMSE versus K, the graph is shown below:

图片包含 地图

描述已自动生成

This graph shows when k = 11(approximately), the RMSE will be minimized, but every time the K which minimize the RMSE maybe different , because each time will have different group of trained and test group. And at this time we run the model, the KNN is 60771.18 as the graph below shows.



And this number is much bigger than the “hand build linear model” which has the RMSE of 58484 (both models using the same trained and test data). This means linear model can predict more accurately than KNN model, so hand build linear model can turn into better performing than KNN model.

So for local taxing authority, they need to levy tax according to the price of house, as the above illustrated, the hand build model has more accurately prediction, so it is better for taxing authority using the hand build linear model to predict price:

**price=bedrooms+ newConstruction+ heating+ fireplaces+ livingArea+ age+waterfront+centralAir+rooms\*heating+landValue\*lotSize+rooms\*bathrooms.**

And according to the predicted price, it can know how much local taxing authority should tax.