Factors Influencing Seattle Residential Housing Prices

Axel Li ECON 484 SP22

Abstract

- Knowing characteristics of a house's effect on housing price in a given year can be helpful for buyers and sellers as a guidance.
- The dataset of previous records was wrangled and various models were built for predicting price, models did not work well.
- The significant change in the ceteris paribus effect of the SqFt of living might lead to the poor performance of models, might result from increased demand

Business Case

Buying a house is a topic that can hardly be avoided for a large number of individuals. How are they priced? It is believed that there should be references of a house's price with given characteristics such as environment, position, and Sqft of living space. I aim to build a model based on previous records of housing exchanges to predict housing prices in 2022. A reliable model can be helpful for us to estimate the housing value with marked price, seeing if there is a good deal or it is overpriced. This can potentially be a guide for people to determine whether to buy, when to buy, where to buy a house.

Hypothesis and Research Questions

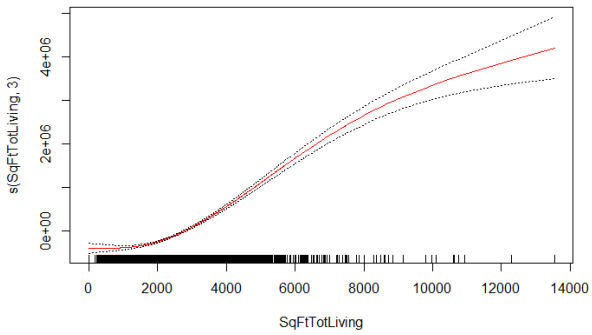
I believe that the housing price should vary among different months in a year, with different districts in terms of positions. I aim to find these patterns through building models. I also want to find out if the market was impacted by the pandemic, that is, are houses becoming more expensive post-covid?

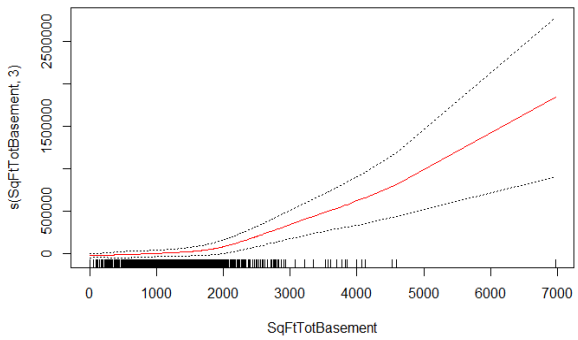
Analysis

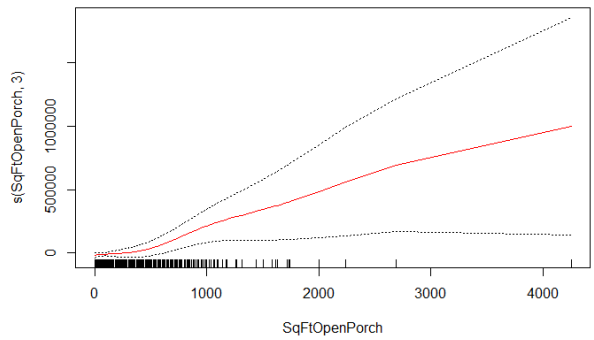
Four datasets were used in the analysis: residential building status, parcel, environmental restrictions of houses, and real property sale history. I combined datasets based on their corresponding identification number: Major and Minor. I manually picked some seemingly relevant variables to analyze: the position, size of attachment, natural environment, ect. Some data wrangling and cleaning were done to exclude not useful records. The cleaned dataset was splitted into the training set: records in 2020 and 2021; and the testing set: records in 2022.

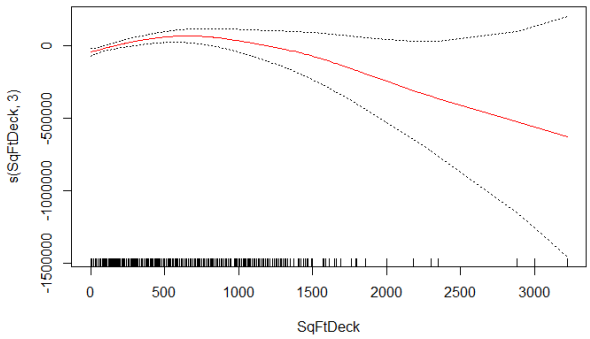
A Linear Regression was performed for intuition check (m1). m1 has an adjusted R-squared of 0.4592. It was then used to fit the test set and a mean squared error of 9.476e+11 was generated.

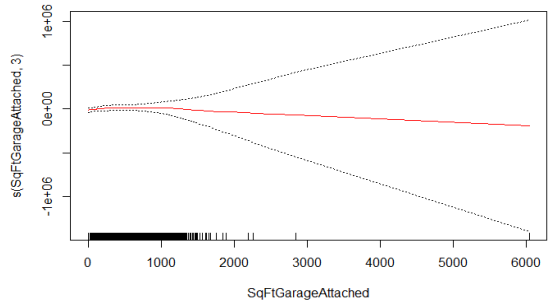
A Generated Additive Model was performed on numerical variables in the training set to see if quadratic relationships exist between sale price and other numeric variables. Plots for some numeric variables in the model are shown below. They had shown whether the underlying relationships were linear or quadratic, and how much influence on SalePrice they led to.

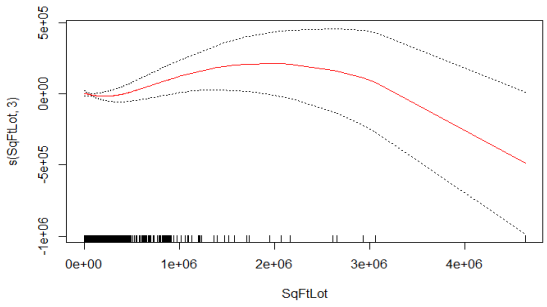












Then I performed a Lasso regression on the training set and used it to fit the testing set . A mean squared error of 8.791e+11 was generated.

According to the result Lasso and GAM shown. I built another Linear Model(m2) with quadratic terms and some variables thrown out. m2 has an adjusted R-squared of 0.4596. It was then used to fit the test set and a mean squared error of 9.474e+11 was generated. A remarkable improvement on performance was not shown.

Despite that Lasso gave the best performance, I argue that the model was not actually useful in terms of prediction. However, I believe it showed us some insights. Based on the Lasso result. SqFt of total living area was the variable with the most explanatory power. This was further supported by an univariate regression(m3) with Rsquared of 0.2592.

Therefore, we performed first-order Double Lasso for the ceteris paribus effect of SqFt of total living area on housing price of 2020, 2021 and 2022 respectively. We found the effect was: 371.5, 377.4, and 652.1 respectively. We observed a huge difference between the training and testing set, the price seemed to be much higher in 2022.

This claim was supported by the fact that among all 833 observations in the testing set, we got 545 occurrences(65.4%) of underprediction, suggesting that the model we had tended to underpredict the price.

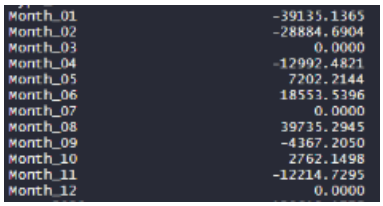
Result and Discussion

Despite the failure of constructing a highly reliable model for price prediction. We still have drawn some insights according to the lasso regression. The analysis process still provided solutions to the question: What characteristics had how much impact on housing prices.

- Houses in Medina, Mercer Island, and Bellevue tend to be more expensive. These districts are right next to each other, potentially due to the well known fact of a better environment. Lasso Results:

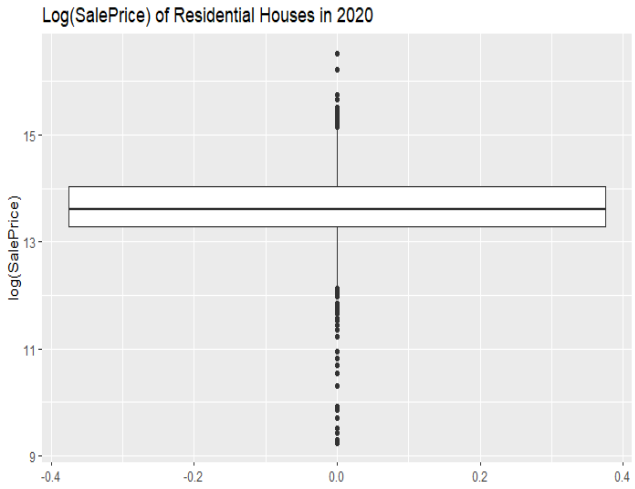
DistrictName_ALGONA	-3368.9369
DistrictName_AUBURN	-21710.8202
DistrictName_BELLEVUE	133330.3289
DistrictName_BLACK DIAMOND	-13960.3441
DistrictName_BOTHELL	4376.1267
DistrictName_BURIEN	0.0000
DistrictName_CARNATION	-14218.3767
DistrictName_CLYDE HILL	47930.2491
DistrictName_COVINGTON	-13588.2205
DistrictName_DES MOINES	-21106.9392
DistrictName_DUVALL	-9320.4125
DistrictName_ENUMCLAW	-5536.3562
DistrictName_FEDERAL WAY	25418.2495
DistrictName_ISSAQUAH	51001.3531
DistrictName_KENMORE	0.0000
DistrictName_KENT	8573.9552
DistrictName_KING COUNTY	-13323.6571
DistrictName_KIRKLAND	30767.2666
DistrictName_LAKE FOREST PARK	-5696.6987
DistrictName_MAPLE VALLEY	-1644.4073
DistrictName_MEDINA	246482.8848
DistrictName_MERCER ISLAND	264258.8936
DistrictName_MILTON	-14806.3750
DistrictName_NEWCASTLE	0.0000
DistrictName_NORMANDY PARK	14871.7392
DistrictName_NORTH BEND	-28826.9135
DistrictName_PACIFIC	-25640.2960
DistrictName_REDMOND	39311.6601
DistrictName_RENTON	-7631.7246
DistrictName_SAMMAMISH	163723.1539
DistrictName_SeaTac	-11012.4001
DistrictName_SEATTLE	59279.3783
DistrictName_SHORELINE	7135.1772
DistrictName_SKYKOMISH	-26059.7634
DistrictName_SNOQUALMIE	-38787.6345
DistrictName_TUKWILA	-1457.7542
DistrictName_WOODINVILLE	0.0000

- Houses tend to be more expensive in August, right before the starting of a school year.

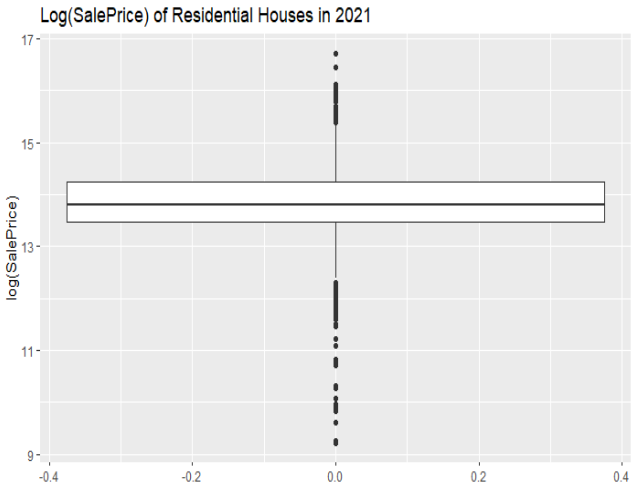


Month_01	-39135.1365
Month_02	-28884.6904
Month_03	0.0000
Month_04	-12992.4821
Month_05	7202.2144
Month_06	18553.5396
Month_07	0.0000
Month_08	39735.2945
Month_09	-4367.2050
Month_10	2762.1498
Month_11	-12214.7295
Month_12	0.0000

- Houses in 2020 tend to be cheaper than 2021. However, this might not be the pure result due to the natural rate of inflation or the continuation of Covid-19. According to the visualizations below, we might not start with a balanced sample in terms of SalePrice.



[mean: 13.61 in 2020]



[mean: 13.81 in 2021]

Inflation had been a serious problem since the money supply was significantly increased as a result from covid relief funds and decreased interest rate. Such a phenomenon was also supported by the skyrocketing causal effect of per sqft of living area, as demand went up since more money was in the market and interest rates went down.
While the test set of 2022 may be considered a post-covid year, we used covid year 2020 and 2021 as the training set. We already expected the condition in the housing market should have been different, and our model had supported such a hypothesis.

Summary Statistics

	2020	2021	2022
Causal Effect of SqftTotLiving on SalePrice	371.5	377.4	652.1
Mean log(SalePrice)	13.61	13.81	13.98

\Model Feature\	M1 (all linear terms)	Lasso	M2 (OLS with adjustments from lasso and GAM)
adj.R2	0.4592	/	0.4596
MSE on test set	9.476e+11	8.791e+11	9.474e+11

Future Direction

A quadratic Double Lasso was not performed due to it being time-consuming, while it could have been a better measurement for the causal effect of Sqft of living area for we should not assume it's independent from other variables.
We did not build a reliable model based on the training set, in terms of R-squared. We could pick more/better variables in the regression through variable selection methods such as forward or backward. Such methods needed a huge amount of preparation on the dataset and I chose not to do so.

Reference

Code were partly adapted from: Maksym Muntyan (Lasso)

Code Appendix:

```
set.seed(484)
library(dplyr)
library(stringr)
library(fastDummies)
library(MLmetrics)
library(glmnet)
library(hdm)

res <- read.csv("EXTR_ResBldg.csv")
parcel <- read.csv("EXTR_Parcel.csv")
env <- read.csv("EXTR_EnvironmentalRestriction_V.csv")
sale <- read.csv("EXTR_RPSale.csv")

# put them together
data_cleaned <- merge(res, env, by = c("Minor" = "Minor", "Major" = "Major"))
data_cleaned <- merge(parcel, data_cleaned, by = c("Minor" = "Minor", "Major" = "Major"))
data_cleaned <- merge(sale, data_cleaned, by = c("Minor" = "Minor", "Major" = "Major"))
# make a column for year and month
data_cleaned$DocumentDate <- as.Date(data_cleaned$DocumentDate, "%m/%d/%Y")
data_cleaned <- data_cleaned %>% mutate(Month = format(data_cleaned$DocumentDate, "%m")) %>%
mutate(Year = format(data_cleaned$DocumentDate,"%Y"))
# do not want empty type
data_cleaned <- data_cleaned %>% filter(Type != "")

# pick variables
data_selected <- data_cleaned %>% dplyr::select(SalePrice, DistrictName, Type, SqFtTotLiving, SqFtLot,
SqFtTotBasement, SqFtOpenPorch, SqFtEnclosedPorch, SqFtDeck, SqFtGarageAttached, PcntUnusable, Month,
Year, BrickStone, HeatSystem, Condition, WaterSystem, SewerSystem, TrafficNoise,  PowerLines, OtherNuisances,
HistoricSite, NbrLivingUnits, BathFullCount)
# We don't want potentially poorly recorded prices(tends out this is really important)
data_selected <- data_selected %>% filter(SalePrice > 10000)
# we are not using all records
data_selected <- data_selected %>% filter(Year > 2019)


# make dummy variables
```

```
data_selected <- dummy_cols(data_selected, select_columns = c("DistrictName", "Type", "Month", "Year"))
data_selected$HeatSystem <- as.factor(data_selected$HeatSystem)
data_selected$Condition <- as.factor(data_selected$Condition)
data_selected$WaterSystem <- as.factor(data_selected$WaterSystem)
data_selected$SewerSystem <- as.factor(data_selected$SewerSystem)
data_selected$TrafficNoise <- as.factor(data_selected$TrafficNoise)
data_selected$PowerLines <- as.factor(data_selected$PowerLines)
data_selected$OtherNuisances <- as.factor(data_selected$OtherNuisances)
data_selected$HistoricSite <- as.factor(data_selected$HistoricSite)
# train test split
train <- data_selected %>% filter(Year < 2022)
test <- data_selected[data_selected$Year == 2022,]
# take out the above non-dummy columns
col_dont_want <- c("DistrictName", "Type", "Month", "Year")
train <- train[, ! names(train) %in% col_dont_want]
test <- test[, ! names(test) %in% col_dont_want]

# Intuitive starting point: a first order lm
m1 <- lm(SalePrice ~ ., data = train)
summary(m1)
m1_pred <- predict(m1, test[,-1])
MSEm1 <- MSE(m1_pred, test$SalePrice)
MSEm1

# Lasso
x <- scale(data.matrix(train[,-1]))
y <- train$SalePrice
cv_model <- cv.glmnet(x, y, alpha = 1)
best_lambda <- cv_model$lambda.min
best_lasso <- glmnet(x, y, alpha = 1, lambda = best_lambda)
as.table(as.matrix(best_lasso$beta))

# Visualize log(price)
train %>% filter(Year_2020 == 1) %>%
  ggplot(aes(y = log(SalePrice))) +
  geom_boxplot() +
  labs(title = "Log(SalePrice) of Residential Houses in 2020")
train %>% filter(Year_2021 == 1) %>%
  ggplot(aes(y = log(SalePrice))) +
  geom_boxplot() +
  labs(title = "Log(SalePrice) of Residential Houses in 2021")


# lasso prediction
x2 <- scale(data.matrix(test[,-1]))
y2 <- test$SalePrice
x2[is.na(x2)] <- 0
lasso_pred <- predict(best_lasso, x2)
MSELasso

# GAM
library(gam)
gam1 <- gam(SalePrice ~ s(SqFtTotLiving,3) + s(SqFtLot,3) + s(SqFtTotBasement,3) + s(SqFtOpenPorch,3) +
s(SqFtEnclosedPorch,3) + s(SqFtDeck,3) + s(SqFtGarageAttached,3) + s(PcntUnusable,3), data = train)
```

```
plot(gam1, col = "red", se = T)

# lm taking account of lasso results and GAM plots
m2 <- lm(SalePrice ~ . - Type_Wetland - Month_12 - Month_07 - Year_2021 - Type_SeismicHazard -
DistrictName_KENMORE - DistrictName_BURIEN - SqFtTotBasement + SqFtLot^2 + SqFtTotLiving^2, data = train)
summary(m2)
m2_pred <- predict(m2, test[,-1])
MSEm2 <- MSE(m2_pred, test$SalePrice)
MSEm2

# Is SqFtTotLiving important?
m3 <- lm(SalePrice ~ SqFtTotLiving, data = train)
summary(m3)$adj.r.squared

# double lasso 2020
fm <- as.formula(~ . - SalePrice - 1)
X <- model.matrix(fm, data = train[train$Year_2020 == 1,])
Y <- train[train$Year_2020 == 1,]$SalePrice
# only want SqFtTotLiving related variables
index.liv <- grep("SqFtTotLiving", colnames(X))
reg.out<-lm(Y~X)
#coefficients for SqFtTotLiving&interactions
index.liv.regout<-grep("SqFtTotLiving",names(reg.out$coefficients))

# Partialling out
# double lasso regress Y on X with the columns in index.liv as focal
effects.liv.ds <- rlassoEffects(x = X, y = Y, method = "double selection",index = index.liv)
summary(effects.liv.ds)

# double lasso 2021
fm <- as.formula(~ . - SalePrice - 1)
X <- model.matrix(fm, data = train[train$Year_2021 == 1,])
Y <- train[train$Year_2021 == 1,]$SalePrice
# only want SqFtTotLiving related variables
index.liv <- grep("SqFtTotLiving", colnames(X))
reg.out<-lm(Y~X)
#coefficients for SqFtTotLiving&interactions
index.liv.regout<-grep("SqFtTotLiving",names(reg.out$coefficients))

# Partialling out
# double lasso regress Y on X with the columns in index.liv as focal
effects.liv.ds <- rlassoEffects(x = X, y = Y, method = "double selection",index = index.liv)
summary(effects.liv.ds)

# double lasso 2022
fm <- as.formula(~ . - SalePrice - 1)
X <- model.matrix(fm, data = test)
Y <- test$SalePrice
# only want SqFtTotLiving related variables
index.liv <- grep("SqFtTotLiving", colnames(X))
reg.out<-lm(Y~X)
#coefficients for SqFtTotLiving&interactions
index.liv.regout<-grep("SqFtTotLiving",names(reg.out$coefficients))
```

```
# Partialling out
# double lasso regress Y on X with the columns in index.liv as focal
effects.liv.ds <- rlassoEffects(x = X, y = Y, method = "double selection",index = index.liv)
summary(effects.liv.ds)

sum(lasso_pred - test$SalePrice < 1)
```