Regression Techniques on House Prices dataset

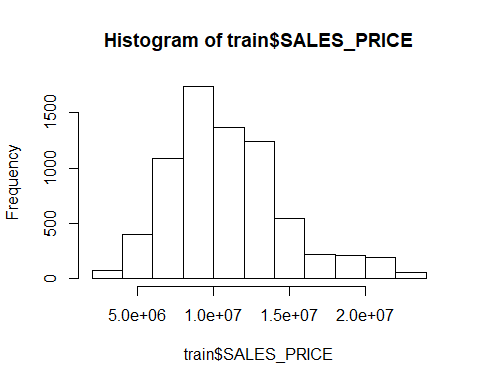
Abhay Padda

10 February 2018

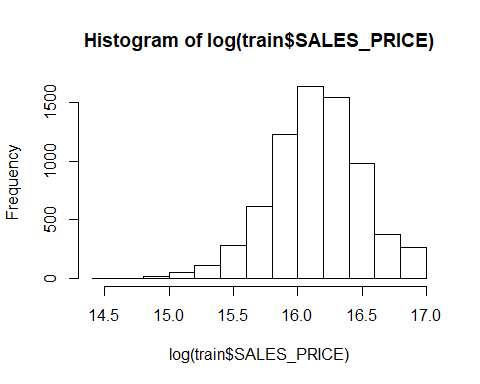
library(zoo)  
library(lubridate)  
  
head(train)

## PRT\_ID AREA INT\_SQFT DATE\_SALE DIST\_MAINROAD N\_BEDROOM N\_BATHROOM  
## 1 P03210 Karapakkam 1004 04-05-2011 131 1 1  
## 2 P09411 AnnaNagar 1986 19-12-2006 26 2 1  
## 3 P01812 Adyar 909 04-02-2012 70 1 1  
## 4 P05346 Velachery 1855 13-03-2010 14 3 2  
## 5 P06210 Karapakkam 1226 05-10-2009 84 1 1  
## 6 P00219 Chrompet 1220 11-09-2014 36 2 1  
## N\_ROOM SALE\_COND PARK\_FACIL DATE\_BUILD BUILDTYPE UTILITY\_AVAIL STREET  
## 1 3 AbNormal Yes 15-05-1967 Commercial AllPub Paved  
## 2 5 AbNormal No 22-12-1995 Commercial AllPub Gravel  
## 3 3 AbNormal Yes 09-02-1992 Commercial ELO Gravel  
## 4 5 Family No 18-03-1988 Others NoSewr Paved  
## 5 3 AbNormal Yes 13-10-1979 Others AllPub Gravel  
## 6 4 Partial No 12-09-2009 Commercial NoSeWa NoAccess  
## MZZONE QS\_ROOMS QS\_BATHROOM QS\_BEDROOM QS\_OVERALL REG\_FEE COMMIS  
## 1 A 4.0 3.9 4.9 4.330 380000 144400  
## 2 RH 4.9 4.2 2.5 3.765 760122 304049  
## 3 RL 4.1 3.8 2.2 3.090 421094 92114  
## 4 I 4.7 3.9 3.6 4.010 356321 77042  
## 5 C 3.0 2.5 4.1 3.290 237000 74063  
## 6 RH 4.5 2.6 3.1 3.320 409027 198316  
## SALES\_PRICE  
## 1 7600000  
## 2 21717770  
## 3 13159200  
## 4 9630290  
## 5 7406250  
## 6 12394750

cleanData <- function(data, trainData=TRUE) {  
 if (trainData) {  
 data$SALES\_PRICE <- log(train$SALES\_PRICE)  
 }  
   
}  
  
hist(train$SALES\_PRICE)



hist(log(train$SALES\_PRICE))



train$SALES\_PRICE <- log(train$SALES\_PRICE)  
  
train$N\_ROOM <- as.factor(train$N\_ROOM)  
train$N\_BATHROOM <- as.factor(train$N\_BATHROOM)  
train$N\_BEDROOM <- as.factor(train$N\_BEDROOM)  
  
train$PROPERTY\_AGE <- (as.yearmon(strptime("11.01.2018", format = "%d.%m.%Y"))-  
 as.yearmon(strptime(train$DATE\_BUILD, format = "%d-%m-%Y")))  
train$SALE\_YEAR <- as.factor(substring(train$DATE\_SALE,7,10))  
  
cleaned\_data <- subset(train, select = -c(PRT\_ID, DATE\_BUILD, DATE\_SALE))  
  
feature\_classes <- sapply(names(cleaned\_data),function(x){class(cleaned\_data[[x]])})  
numeric\_feats <-names(feature\_classes[feature\_classes != "factor"])   
  
library(moments)  
skewed\_feats <- sapply(numeric\_feats,function(x){skewness(cleaned\_data[[x]],na.rm=TRUE)})  
  
skewed\_feats <- sapply(numeric\_feats,function(x){skewness(cleaned\_data[[x]],na.rm=TRUE)})  
skewed\_feats <- skewed\_feats[skewed\_feats > 0.75]  
  
for(x in names(skewed\_feats)) {  
 cleaned\_data[[x]] <- log(cleaned\_data[[x]] + 1)  
}  
  
categorical\_feats <- names(feature\_classes[feature\_classes == "factor"])  
  
sapply(cleaned\_data, function(x) sum(is.na(x)))

## AREA INT\_SQFT DIST\_MAINROAD N\_BEDROOM N\_BATHROOM   
## 0 0 0 1 5   
## N\_ROOM SALE\_COND PARK\_FACIL BUILDTYPE UTILITY\_AVAIL   
## 0 0 0 0 0   
## STREET MZZONE QS\_ROOMS QS\_BATHROOM QS\_BEDROOM   
## 0 0 0 0 0   
## QS\_OVERALL REG\_FEE COMMIS SALES\_PRICE PROPERTY\_AGE   
## 48 0 0 0 0   
## SALE\_YEAR   
## 0

Mode <- function (x, na.rm) {  
 xtab <- table(x)  
 xmode <- names(which(xtab == max(xtab)))  
 if (length(xmode) > 1) xmode <- ">1 mode"  
 return(xmode)  
}  
  
for (var in 1:ncol(cleaned\_data)) {  
 if (class(cleaned\_data[,var])=="numeric") {  
 cleaned\_data[is.na(cleaned\_data[,var]),var] <- mean(cleaned\_data[,var], na.rm = TRUE)  
 } else if (class(cleaned\_data[,var]) %in% c("character", "factor")) {  
 cleaned\_data[is.na(cleaned\_data[,var]),var] <- Mode(cleaned\_data[,var], na.rm = TRUE)  
 }  
}  
  
sapply(cleaned\_data, function(x) sum(is.na(x)))

## AREA INT\_SQFT DIST\_MAINROAD N\_BEDROOM N\_BATHROOM   
## 0 0 0 0 0   
## N\_ROOM SALE\_COND PARK\_FACIL BUILDTYPE UTILITY\_AVAIL   
## 0 0 0 0 0   
## STREET MZZONE QS\_ROOMS QS\_BATHROOM QS\_BEDROOM   
## 0 0 0 0 0   
## QS\_OVERALL REG\_FEE COMMIS SALES\_PRICE PROPERTY\_AGE   
## 0 0 0 0 0   
## SALE\_YEAR   
## 0

library(caret)  
dummies <- dummyVars(~.,cleaned\_data[categorical\_feats])  
categorical\_1\_hot <- predict(dummies,cleaned\_data[categorical\_feats])  
  
final <- cbind(cleaned\_data[numeric\_feats],categorical\_1\_hot)

x<- subset(final,select= -SALES\_PRICE)  
y <- final$SALES\_PRICE

CARET.TRAIN.CTRL <- trainControl(method="repeatedcv", number=5, repeats=5, returnResamp="final", verboseIter=FALSE)  
  
## LINEAR REGRESSION  
model\_linear <- train(SALES\_PRICE~.,final,method="lm",metric="RMSE",maximize=FALSE,trControl=CARET.TRAIN.CTRL)  
summary(model\_linear)

##   
## Call:  
## lm(formula = .outcome ~ ., data = dat)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.38396 -0.02400 -0.00028 0.02213 0.20143   
##   
## Coefficients: (13 not defined because of singularities)  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.404e+01 4.975e-02 282.294 < 2e-16 \*\*\*  
## INT\_SQFT 3.734e-04 5.126e-06 72.858 < 2e-16 \*\*\*  
## DIST\_MAINROAD -2.319e-06 8.852e-06 -0.262 0.793349   
## QS\_ROOMS -4.573e-03 1.563e-03 -2.927 0.003435 \*\*   
## QS\_BATHROOM -4.572e-03 1.674e-03 -2.731 0.006328 \*\*   
## QS\_BEDROOM -3.926e-03 1.909e-03 -2.056 0.039813 \*   
## QS\_OVERALL 2.034e-02 4.877e-03 4.172 3.06e-05 \*\*\*  
## REG\_FEE 1.129e-01 3.828e-03 29.499 < 2e-16 \*\*\*  
## COMMIS 8.361e-03 1.111e-03 7.525 5.94e-14 \*\*\*  
## PROPERTY\_AGE -1.758e-03 4.418e-05 -39.788 < 2e-16 \*\*\*  
## AREA.Adyar 1.554e-01 4.119e-03 37.736 < 2e-16 \*\*\*  
## AREA.AnnaNagar 1.845e-01 3.269e-03 56.457 < 2e-16 \*\*\*  
## AREA.Chrompet 1.293e-01 3.609e-03 35.839 < 2e-16 \*\*\*  
## AREA.Karapakkam -1.213e-01 3.742e-03 -32.414 < 2e-16 \*\*\*  
## AREA.KKNagar -6.847e-02 2.565e-03 -26.691 < 2e-16 \*\*\*  
## AREA.TNagar 1.884e-01 3.404e-03 55.342 < 2e-16 \*\*\*  
## AREA.Velachery NA NA NA NA   
## N\_BEDROOM.1 8.426e-02 6.698e-03 12.579 < 2e-16 \*\*\*  
## N\_BEDROOM.2 8.091e-02 5.504e-03 14.701 < 2e-16 \*\*\*  
## N\_BEDROOM.3 5.269e-02 3.900e-03 13.512 < 2e-16 \*\*\*  
## N\_BEDROOM.4 NA NA NA NA   
## N\_BATHROOM.1 -5.760e-03 3.472e-03 -1.659 0.097162 .   
## N\_BATHROOM.2 NA NA NA NA   
## N\_ROOM.2 -7.010e-02 4.166e-03 -16.825 < 2e-16 \*\*\*  
## N\_ROOM.3 -1.916e-02 3.571e-03 -5.367 8.28e-08 \*\*\*  
## N\_ROOM.4 NA NA NA NA   
## N\_ROOM.5 NA NA NA NA   
## N\_ROOM.6 NA NA NA NA   
## SALE\_COND.AbNormal 2.444e-02 1.613e-03 15.150 < 2e-16 \*\*\*  
## SALE\_COND.AdjLand 5.578e-02 1.620e-03 34.439 < 2e-16 \*\*\*  
## SALE\_COND.Family 1.548e-02 1.614e-03 9.589 < 2e-16 \*\*\*  
## SALE\_COND.NormalSale 2.765e-02 1.610e-03 17.179 < 2e-16 \*\*\*  
## SALE\_COND.Partial NA NA NA NA   
## PARK\_FACIL.No -8.995e-02 1.105e-03 -81.400 < 2e-16 \*\*\*  
## PARK\_FACIL.Yes NA NA NA NA   
## BUILDTYPE.Commercial 2.909e-01 1.824e-03 159.522 < 2e-16 \*\*\*  
## BUILDTYPE.House -6.385e-02 1.274e-03 -50.135 < 2e-16 \*\*\*  
## BUILDTYPE.Others NA NA NA NA   
## UTILITY\_AVAIL.AllPub 1.100e-02 1.407e-03 7.822 5.96e-15 \*\*\*  
## UTILITY\_AVAIL.ELO -1.752e-02 1.512e-03 -11.585 < 2e-16 \*\*\*  
## UTILITY\_AVAIL.NoSeWa -1.160e-02 1.409e-03 -8.238 < 2e-16 \*\*\*  
## UTILITY\_AVAIL.NoSewr NA NA NA NA   
## STREET.Gravel 4.560e-02 1.218e-03 37.432 < 2e-16 \*\*\*  
## STREET.NoAccess -5.915e-02 1.340e-03 -44.136 < 2e-16 \*\*\*  
## STREET.Paved NA NA NA NA   
## MZZONE.A -3.193e-01 2.703e-03 -118.111 < 2e-16 \*\*\*  
## MZZONE.C -2.427e-01 2.509e-03 -96.741 < 2e-16 \*\*\*  
## MZZONE.I -1.658e-01 2.467e-03 -67.220 < 2e-16 \*\*\*  
## MZZONE.RH -9.931e-02 1.484e-03 -66.914 < 2e-16 \*\*\*  
## MZZONE.RL -4.893e-02 1.429e-03 -34.244 < 2e-16 \*\*\*  
## MZZONE.RM NA NA NA NA   
## SALE\_YEAR.2004 2.787e-02 7.446e-03 3.743 0.000183 \*\*\*  
## SALE\_YEAR.2005 1.844e-02 7.522e-03 2.451 0.014255 \*   
## SALE\_YEAR.2006 2.032e-02 6.677e-03 3.044 0.002345 \*\*   
## SALE\_YEAR.2007 1.701e-02 6.254e-03 2.720 0.006535 \*\*   
## SALE\_YEAR.2008 1.663e-02 6.222e-03 2.672 0.007549 \*\*   
## SALE\_YEAR.2009 1.428e-02 6.179e-03 2.310 0.020895 \*   
## SALE\_YEAR.2010 1.266e-02 6.171e-03 2.051 0.040319 \*   
## SALE\_YEAR.2011 1.114e-02 6.173e-03 1.804 0.071254 .   
## SALE\_YEAR.2012 1.053e-02 6.245e-03 1.686 0.091928 .   
## SALE\_YEAR.2013 7.284e-03 6.398e-03 1.138 0.254960   
## SALE\_YEAR.2014 2.755e-03 6.437e-03 0.428 0.668698   
## SALE\_YEAR.2015 NA NA NA NA   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.04269 on 7059 degrees of freedom  
## Multiple R-squared: 0.9853, Adjusted R-squared: 0.9852   
## F-statistic: 9639 on 49 and 7059 DF, p-value: < 2.2e-16

mean(model\_linear$resample$RMSE)

## [1] 0.0428548

## RIDGE REGRESSION  
set.seed(123) # for reproducibility  
model\_ridge <- train(x=x,y=y, method="glmnet", metric="RMSE",maximize=FALSE,trControl=CARET.TRAIN.CTRL,tuneGrid=expand.grid(alpha=0, lambda=0.039)) #alpha is set to 0 for Ridge regression

## Loading required package: glmnet

## Loading required package: Matrix

## Loading required package: foreach

## Loaded glmnet 2.0-10

mean(model\_ridge$resample$RMSE)

## [1] 0.05149145

## LASSO   
set.seed(123) # for reproducibility  
model\_lasso <- train(x=x,y=y,  
method="glmnet",  
metric="RMSE",  
maximize=FALSE,  
trControl=CARET.TRAIN.CTRL,  
tuneGrid=expand.grid(alpha=1,lambda=0.01)) # alpha is set to 1 for Lasso regression  
  
model\_lasso

## glmnet   
##   
## 7109 samples  
## 62 predictor  
##   
## No pre-processing  
## Resampling: Cross-Validated (5 fold, repeated 5 times)   
## Summary of sample sizes: 5687, 5686, 5687, 5687, 5689, 5686, ...   
## Resampling results:  
##   
## RMSE Rsquared   
## 0.07106875 0.9622713  
##   
## Tuning parameter 'alpha' was held constant at a value of 1  
##   
## Tuning parameter 'lambda' was held constant at a value of 0.01  
##

mean(model\_lasso$resample$RMSE)

## [1] 0.07106875