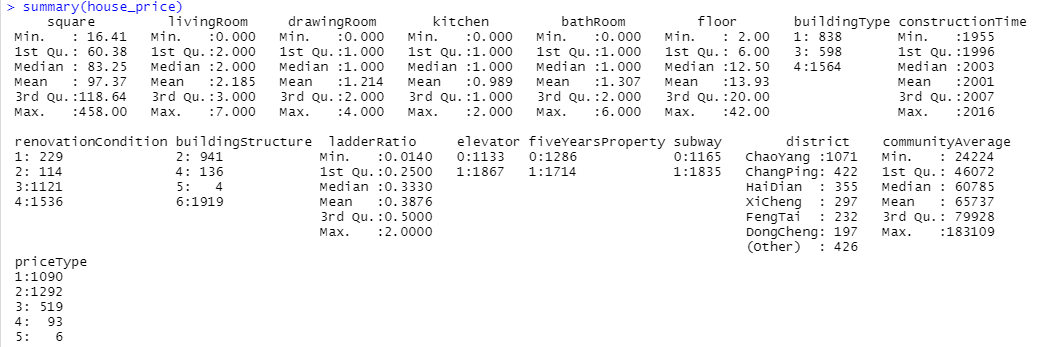
一.数据说明与基本数据处理

由于数据集规模过大，行数接近30万，在跑svm算法时需要时间过长，于是我截取了数据的前3万行进行测试

数据展示：



library(ggplot2)

library(caret)

library(e1071)

house\_price<-read.csv(file.choose())

#概览数据集，并且进行变量类型转换

summary(house\_price)

class(house\_price[,7])

house\_price[,7]<-as.factor(house\_price[,7])

class(house\_price[,9])

house\_price[,9]<-as.factor(house\_price[,9])

class(house\_price[,10])

house\_price[,10]<-as.factor(house\_price[,10])

class(house\_price[,12])

house\_price[,12]<-as.factor(house\_price[,12])

class(house\_price[,13])

house\_price[,13]<-as.factor(house\_price[,13])

class(house\_price[,14])

house\_price[,14]<-as.factor(house\_price[,14])

class(house\_price[,17])

house\_price[,17]<-as.factor(house\_price[,17])

head(house\_price)

#设置随机种子并对数据集进行处理

set.seed(5032)

ind<-sample(2,nrow(house\_price),replace=T,prob=c(0.7,0.3))#divide train-data and test-data

trainData<-house\_price[ind==1,]

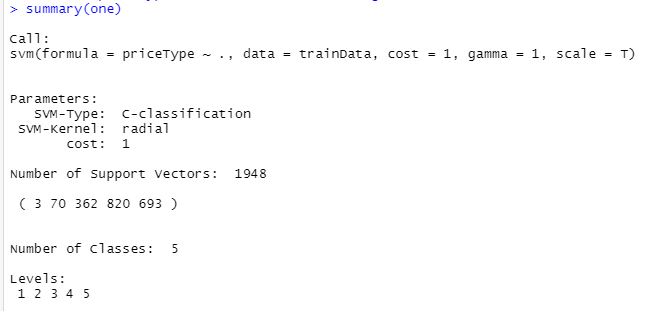
testData<-house\_price[ind==2,]set.seed(125)

二.设置svm模型查看情况

#用一个模型查看情况

one<-svm(priceType~.,data=trainData,cost=1,gamma=1,scale=T)

summary(one)

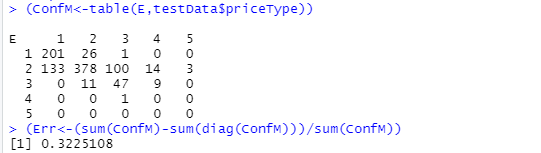


one\_predict<-predict(one,testData)

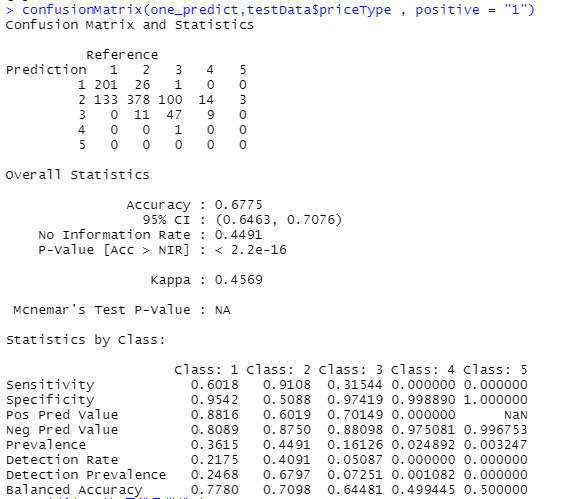
E<-predict(one,testData)

(ConfM<-table(E,testData$priceType))

(Err<-(sum(ConfM)-sum(diag(ConfM)))/sum(ConfM))



confusionMatrix(one\_predict,testData$priceType , positive = "1")



可以看出此时的准确率为0.6775，是偏低的，为了优化模型，利用10折交叉验证找到预测误差最小下的最优参数和最优模型

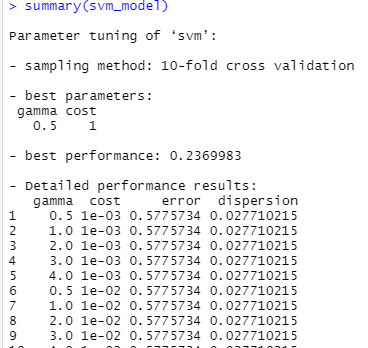
三.优化模型

利用10折交叉验证找到预测误差最小下的最优参数和最优模型，由于该模型用3万条数据跑出来耗费时间过长，所以只挑选了最初数据集中的前3千条数据进行最优模型查找和测试

#用十折交叉验证寻找最优模型

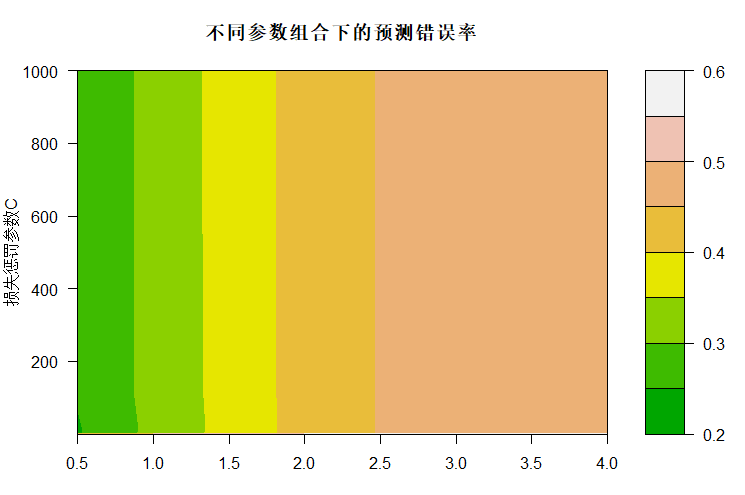
svm\_model<-tune.svm(priceType~.,data=trainData,cost=c(0.001,0.01,0.1,1,5,10,100,1000),gamma=c(0.5,1,2,3,4),scale=T)

summary(svm\_model)



可视化10折交叉验证后的不同参数组合下的结果

plot(svm\_model,xlab=expression(gamma),ylab="损失惩罚参数C",main="不同参数组合下的预测错误率",nlevels=10,color.palette=terrain.colors)

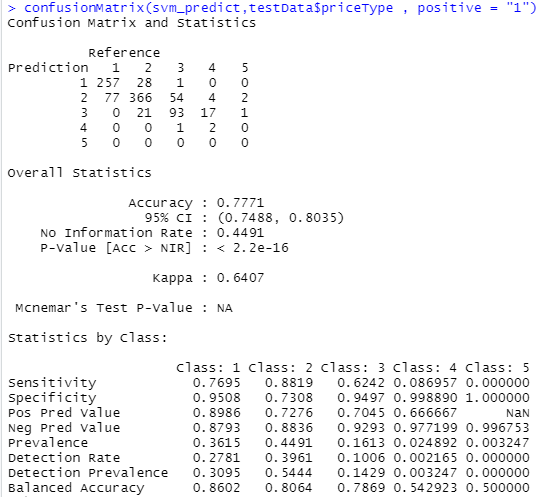


从上图和对tune.svm的summary可看出，模型最优、错误率最低的模型为惩罚因子C=1，gamma=0.5时的模型，用该最优模型进行预测

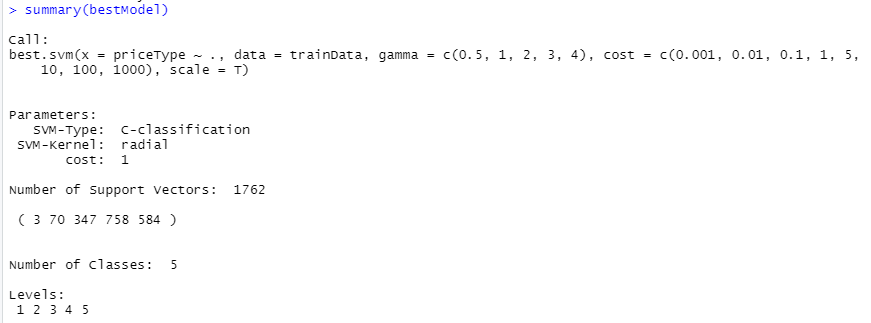
bestModel<-svm\_model$best.model

svm\_predict<-predict(bestModel,testData)

confusionMatrix(svm\_predict,testData$priceType , positive = "1")



summary(bestModel)



此时的准确率达到0.7771，所有包含的支持向量数为1762个，