Predicting Housing Prices in King’s County, WA

Sahil & Abhishek

4/21/2017

**Introduction**

1. Discuss origin and structure of data. What is missing? Why was it collected? How was it collected? What makes it interesting (Think in terms of the dimensions of big data: velocity, volume, variety)

Zillow is the largest online web agency which displays prices of houses all around the United States for property management agencies. Thus Zillow will certainly have a data base of all the housing prices which it displays along with features of the house. Thus, an Analytics consulting agency has tied up with Zillow to receive all this data. The objective of this company is to tie up with other real estate agents to take over the analytics part of the job.

1. Discuss problem being studied in terms of CoNVO.

**Context:**

The data is used by an Analytics consulting company to build models for various real estate clients such that they can accurately display housing prices to their customers.

**Need:**

The Analytics consulting company needs to accurately create models such that the real estate clients are satisfied with the pricing and do not lose out on their customer base.

**Vision:**

We hope to analyze the data to achieve some kind of relationship between the prices of houses with the various factors that affect it. The decision that can result from these outcomes is to find the accurate price for a house with a given set of attributes.

**Outcome:**

The results of this data analysis will be used by the clients on the Analytics Consulting company to display the correct prices for new houses. Success will be measured by increase in number of buyers in Kings County from the past years.

1. What were obstacles you faced while analysing the data? How did you deal with data that you would have wanted but was not available?

Obstacles:

* Time: It is very time consuming to evaluate models like Random Forest and MARS
* Geographical Restriction: This data is only restricted to the Kings County area and the model cannot be generalised for the entire state or more than one county. As a consulting company we will require more data to generalize this for clients.

1. Hypothesis. What is the believed real world relationship between the predictors and outcome (or the counterparts as applicable to your project)

In the real world, predictors like ‘bedrooms’,’sqft\_living’, and ‘yr\_built’ should surely have a linear relationship with the price.

The more the number of bedrooms, the more the rent of a house. This should hold true even for the other predictors like square feet living, and the year built.

1. Analysis - Methodology. How it relates predictors to outcomes

The steps used in the Analysis of the data include:

* Visualizations & Plots
* Data Pre-processing and cleaning
* Exploratory Analysis
* Preliminary Analysis
* Model Evaluation and Prediction

**Model Evaluation**

We had finalized **Random Forest Model** on our last project submission. We are going to perform random forest on training set again and implement the model on test data set and compare the predicted prices with prices mentioned in test data set.

kc\_house\_data <- read.csv("~/Downloads/kc\_house\_data.csv", header=T)  
library(lubridate)

##   
## Attaching package: 'lubridate'

## The following object is masked from 'package:base':  
##   
## date

library(rpart)  
library(caret)

## Loading required package: lattice

## Loading required package: ggplot2

library(ggplot2)  
library(GGally)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following object is masked from 'package:GGally':  
##   
## nasa

## The following objects are masked from 'package:lubridate':  
##   
## intersect, setdiff, union

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

#changing the year of non-renovated house to the year built  
for (i in 1:nrow(kc\_house\_data)) {  
 if (kc\_house\_data$yr\_renovated[i]==0) {  
 kc\_house\_data$yr\_renovated[i]=kc\_house\_data$yr\_built[i]  
 }  
}

#splitting the dataset into 20% test set and 80% training set  
ratio <- sample(1:nrow(kc\_house\_data),size=0.20\*nrow(kc\_house\_data))  
kc\_house\_test <- kc\_house\_data[ratio,]  
kc\_house\_train <- kc\_house\_data[-ratio,]

#Random forest model  
library(randomForest)

## randomForest 4.6-12

## Type rfNews() to see new features/changes/bug fixes.

##   
## Attaching package: 'randomForest'

## The following object is masked from 'package:dplyr':  
##   
## combine

## The following object is masked from 'package:ggplot2':  
##   
## margin

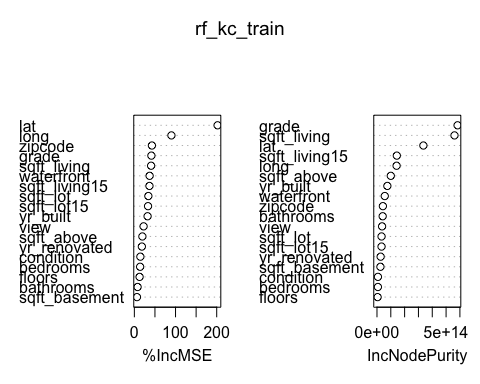
set.seed(7)  
start <- Sys.time()  
rf\_kc\_train <- randomForest(x=kc\_house\_train[,-c(1:3)],y=kc\_house\_train$price,data= kc\_house\_train,mtry=10,importance=T,scale=T)  
end <- Sys.time()  
timecost <- end - start  
print(timecost)

## Time difference of 5.989798 mins

# measuring the importance of predictors on the model  
importance(rf\_kc\_train)

## %IncMSE IncNodePurity  
## bedrooms 13.982806 6.736987e+12  
## bathrooms 7.843824 4.123117e+13  
## sqft\_living 40.674828 5.608506e+14  
## sqft\_lot 33.987402 3.307978e+13  
## floors 12.684912 4.641006e+12  
## waterfront 37.063468 5.637741e+13  
## view 22.415043 3.687943e+13  
## condition 14.575643 6.964471e+12  
## grade 41.341423 5.831729e+14  
## sqft\_above 19.505344 9.994584e+13  
## sqft\_basement 6.369636 2.489135e+13  
## yr\_built 32.126910 7.499326e+13  
## yr\_renovated 18.217001 2.639970e+13  
## zipcode 42.629679 4.276255e+13  
## lat 202.119887 3.357294e+14  
## long 90.009745 1.419464e+14  
## sqft\_living15 36.687036 1.434573e+14  
## sqft\_lot15 33.229662 3.294008e+13

varImpPlot(rf\_kc\_train)



These two plots show the effect of predictors on the model accuracy. If the values of a predictor are randomly permuted, then the accuracy decreases if the predictor has any significance on the model else the model accuracy remains the same. We can see that latitude predictor has the highest significance. This implies that the direction North-South is an important feature for selling the house.

#calculating rmse of the test data set  
predictions <- floor(predict(rf\_kc\_train,kc\_house\_test))  
error=predictions-kc\_house\_test$price  
rmse <- sqrt(mean(error^2))  
rmse

## [1] 128851.3

The RMSE value of the test data set is almost the same as for trained model.

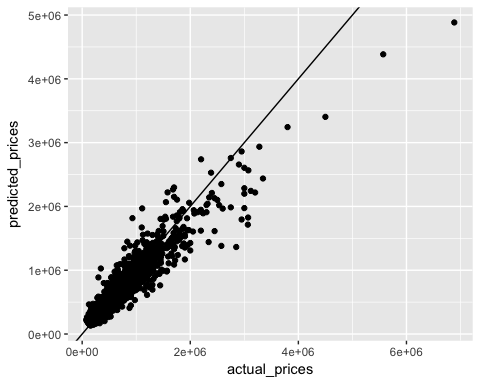
# Comparing predicted prices and actual house prices   
df1 <- data.frame(predictions,kc\_house\_test$price)  
colnames(df1) <- c("predicted\_prices", "actual\_prices")  
  
#Calculating accuracy rate  
error\_percent <- ((abs(df1$predicted\_prices-df1$actual\_prices))/df1$actual\_prices)\*100  
accuracy\_percent <- 100-error\_percent  
df0<- cbind(df1,accuracy\_percent)  
summary(accuracy\_percent)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -171.30 83.40 91.09 86.78 96.26 100.00

# % of predictions when the error was more than the actual house prices  
count=0  
for (i in 1:length(accuracy\_percent)) {  
 if (accuracy\_percent[i]< 0) {  
 count=count+1  
 }  
}  
count\_percent <- (count/length(accuracy\_percent))\*100  
count\_percent

## [1] 0.5090236

# % of predictions when accuracy rate was atleast 70%  
d1 <- df0[-which(accuracy\_percent>=70 & accuracy\_percent<=100),]  
ggplot(data=df0,aes(actual\_prices,predicted\_prices))+ geom\_point()+geom\_abline()



(1-nrow(d1)/nrow(df0))\*100

## [1] 90.6062

**Conclusion**

As you can see from the above plot of predicted prices vs actual prices of the test dataset, most of the points fall around the line which says that most of the predictions were approximately accurate. About 91% of the predictions were at least 70 % accurate. Hence we can safely say that our model is a success and can be implemented for predicting all the new houses for sale in King County. Hence the real estate agents can benefit a lot from this rather than coming up with a price based on their intuition