Analysis of Real Estate Data.

Vamsi Kunaparaju, Aditya Job. **- 14 May 2017.**

Under the Guidance of:

Mr. Ran Ji.

George Mason University.

Background:

Real estate industry is one thing every individual comes across at least once in their lifetime. Everyone needs a place to live. The life ambition of the poor, the most important objective of the middle class and the very interesting aspect for the rich, is to have a place of their own. Buying a house isn't an easy decision to make and, there are many a several factors that are to be considered before buying a house. So, the price of the house must be reasonable both to the seller and the buyer, for the sale to happen sanely. This is the reason that motivated us to pick the data, that is related to the pricing of the houses as per the various factors and commodities.

Problem Description:

Though, there were many places we searched for the data and most of them seemed relevant, this data set we found in KAGGLE seemed really interesting. With 79 explanatory variables describing (almost) every aspect of residential homes in Ames, Iowa, this dataset challenges us to predict the final price of each home. The real challenge of dealing with this dataset is not predicting the house prices but dealing with 79 explanatory variables. Many of these 79 variables turned out to be categorical varying from different levels. So, cleaning the dataset and finding the important variables might turn out to be a very important step when compared to the cleaning procedures of other datasets. The main idea of our project remains to clean the dataset and make it ready for modelling, prediction and find some interesting points on our way through exploratory and statistical analysis.

Data Description:

The dataset was picked from Kaggle website. This playground competition's dataset proves that much more influences price negotiations than the number of bedrooms or a white-picket fence(1). The Ames housing dataset was compiled by Dean De Cock for use in data science education. It's an incredible alternative for data scientists looking for a modernized and expanded version of the often cited Boston Housing dataset(2). Data available for this project contains following data fields,

```
##  [1] "Id"           "MSSubClass"    "MSZoning"      "LotFrontage"
##  [5] "LotArea"      "Street"        "Alley"         "LotShape"
##  [9] "LandContour"  "Utilities"     "LotConfig"     "LandSlope"
## [13] "Neighborhood" "Condition1"    "Condition2"    "BldgType"
## [17] "HouseStyle"   "OverallQual"   "OverallCond"   "YearBuilt"
## [21] "YearRemodAdd" "RoofStyle"     "RoofMatl"      "Exterior1st"
## [25] "Exterior2nd"  "MasVnrType"    "MasVnrArea"    "ExterQual"
## [29] "ExterCond"    "Foundation"    "BsmtQual"      "BsmtCond"
```

(1)- As stated in the website Kaggle for dataset and problem description

(2)- Lines used in Kaggle to mention the origin of the dataset.

```
## [33] "BsmtExposure"   "BsmtFinType1"   "BsmtFinSF1"     "BsmtFinType2"
## [37] "BsmtFinSF2"     "BsmtUnfSF"      "TotalBsmtSF"    "Heating"
## [41] "HeatingQC"      "CentralAir"     "Electrical"     "X1stFlrSF"
## [45] "X2ndFlrSF"      "LowQualFinSF"   "GrLivArea"      "BsmtFullBath"
## [49] "BsmtHalfBath"   "FullBath"       "HalfBath"       "BedroomAbvGr"
## [53] "KitchenAbvGr"   "KitchenQual"    "TotRmsAbvGrd"   "Functional"
## [57] "Fireplaces"     "FireplaceQu"    "GarageType"     "GarageYrBlt"
## [61] "GarageFinish"   "GarageCars"     "GarageArea"     "GarageQual"
## [65] "GarageCond"     "PavedDrive"     "WoodDeckSF"     "OpenPorchSF"
## [69] "EnclosedPorch"  "X3SsnPorch"     "ScreenPorch"    "PoolArea"
## [73] "PoolQC"         "Fence"          "MiscFeature"    "MiscVal"
## [77] "MoSold"         "YrSold"         "SaleType"       "SaleCondition"
## [81] "SalePrice"
```

Visualizations and Inferences:

This phase was particularly chosen to find out some interesting facts and stories that the data had to tell us. We made different plots with variables that we think might have high influence with the target variable i.e. ‘Sale Price’.

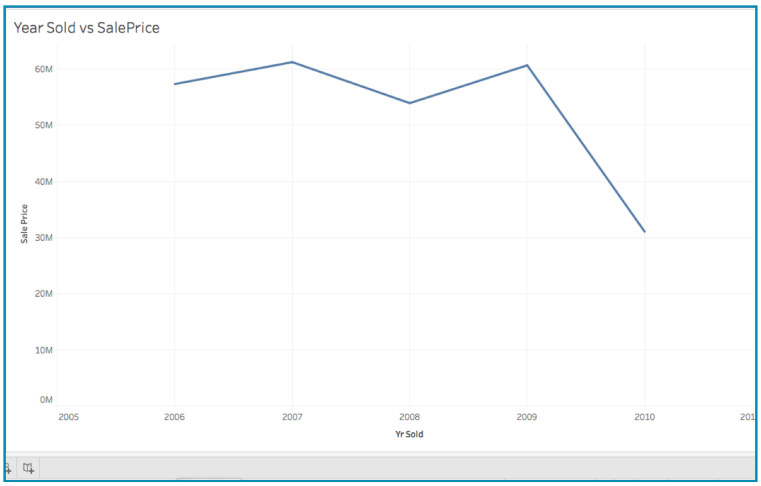


Fig: sale price (vs) year the house has been sold.

Observe the pattern from 2007 to 2009?

- This period is the sub-prime crisis.
- In 2008, the market had hit rock bottom.
- From 2008, the banks were aided with stimulus packages.

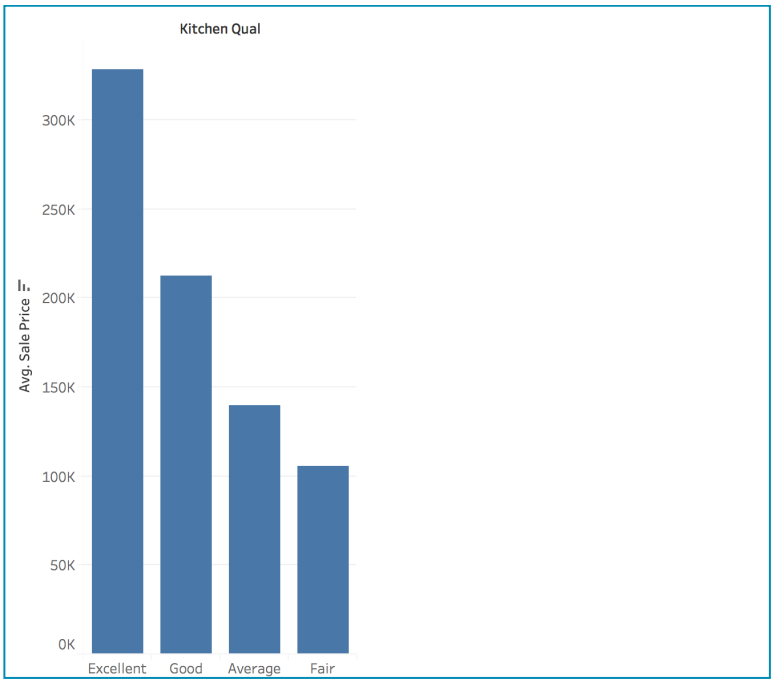


Fig: Sale price (vs) Kitchen Quality

One interesting observation made from the data is that the 'kitchen quality' influences the final sale price by a lot. It might be because of the nature of the American people to use the kitchen and dining as a hangout place. Whereas, people from some other countries do not spend as much time in the kitchen.

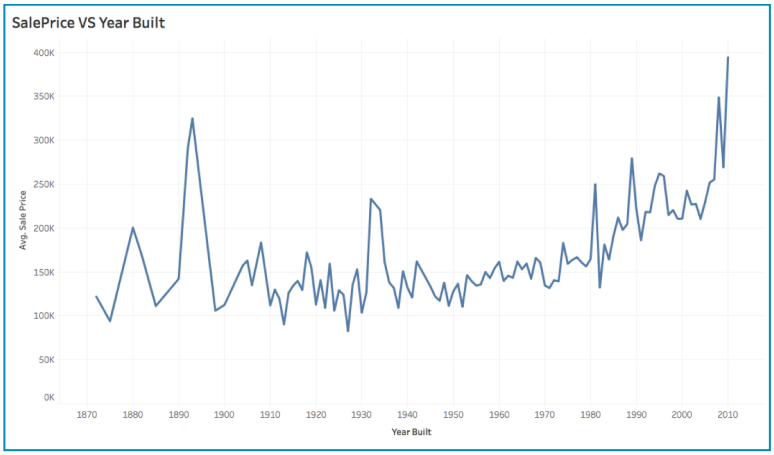


Fig: sale price (vs) Year the house has been built.

The highs and lows in the line above correlates well with the real estate industry's status.

- 2000-2005 Real estate Boom..!

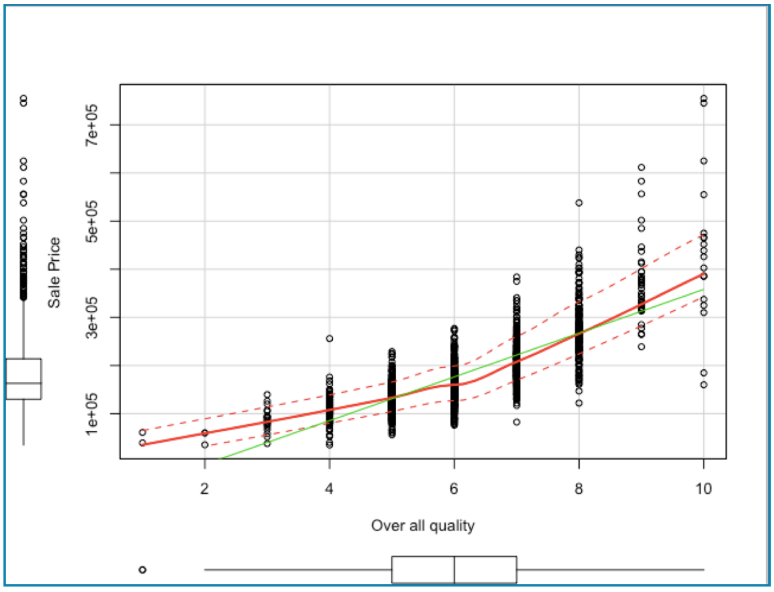


Fig: sale price (vs) Overall quality of the house.

The variable, 'Overall quality' didn't come with much explanation in the dataset description. We assumed it to be a variable that has been calculated considering the values of other variables. It sure had high influence and linearity with our target variable helping the models to predict better.

Data cleaning and Preparation:

This cleaning and preparation has many sub phases and data cleaning has been a major part in processing this dataset.

Data Cleaning: cleaning the data set includes dealing with NA values, dealing with many errors that might occur while collecting the data (e.g. spelling mistakes). Removal of NA values can be dealt in two ways, removing the records that has Na values or imputing the records with the NA values with the mean or average of the whole column. Since, the records that had NA values in our dataset are very less we considered removing them than imputing them with a mean value.

- Declaring all old cleaned predictors "Null".
- Omitting of missing values using "na.omit" function.
- For example,

```
mydata_without_na = na.omit(mydata)
```

Data Transformation: The data thus clean from NA values and error values is ready for transformation. In this phase, we had to convert multiple categorical variables to continuous by assigning each categorical value with a number and establishing a relation between them (e.g. Best-3, good-2, fair-1). Around 20 categorical variables have been converted to continuous using this technique.

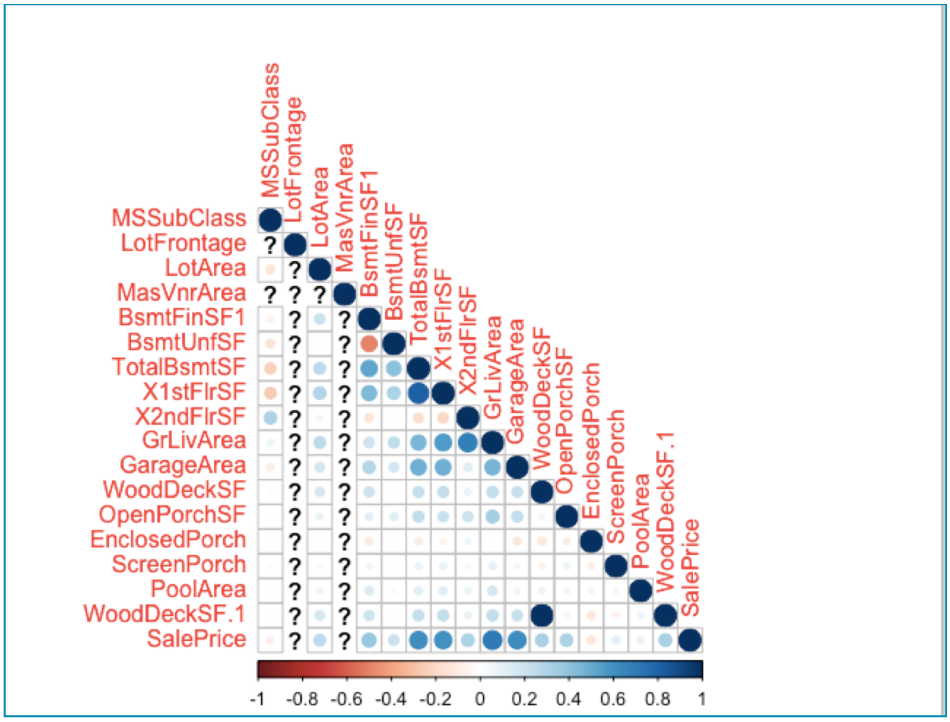
Following is an example, how the variables in the data have been transformed.

```
price <- summarize(group_by(mydata,KitchenQual  ),mean(SalePrice, na.rm=T))
1       Ex              328554.7
2       Fa              105565.2
3       Gd               212116.0
4       TA                              139962.5
mydata$kitchen[mydata$KitchenQual == "Ex"] <- 4
mydata$kitchen[mydata$KitchenQual == "Gd"] <- 3
mydata$kitchen[mydata$KitchenQual == "TA"] <- 2
mydata$kitchen[mydata$KitchenQual == "Fa"] <- 1
```

Data Preparation: The data thus transformed is ready for modelling. But, in order to model the data, it is very important to understand the data. So, lots of exploratory analysis has been done during this phase. It was understood that we need to understand how the relation between various variables exist and how influential they are to each other. In other words, problems like linearity were dealt at this phase.

Multicollinearity

```
       LotArea    housefunction    exterior_cond       WoodDeckSF
      1.248758         1.109540         2.950290         1.152264
       kitchen         PoolArea      OverallCond         bsmt_exp
      2.407375         1.077213         1.128279         1.262944
    MSSubClass       GarageCars       sale_cond nbhd_price_level
      1.244885         1.839565         1.206237         1.860649
   OverallQual        X2ndFlrSF        X1stFlrSF     culdesac_fr3
      3.377490         1.691936         2.184773         1.059281
    MasVnrArea
      1.312804
>
```



On analysing the multicollinearity table, we can see that the VIF values of each predictor is less than ten. These values shows us that these predictors are not having multi-collinearity problem and contribute individually to the prediction of the target variable.

Linear Regression:

Introduction: Linear regression is a statistical procedure used to predict the value of a dependent variable by looking at the independent variable, when, the relationship between the variables can be described with a linear model. It assumes that the data is normally distributed. Though, Linear model isn't the best model to predict the values of a target variable, it can also be considered as a base model with which we can compare the performance of other advanced models.

Data Preparation: To design a linear model, both the independent variable and the target variable must be continuous. The dataset has been normalised and was randomly split into training (60%) and validation (40%) sets.

Model Tuning: First a full model has been run on the data by throwing all the available predictors (not considering any significance) and the value for comparison i.e. root mean squared error has been calculated. Then, the forward and backward regression steps are performed on the full model, which yielded better results than the full model. Finally, a reduced model has been run with the predictors that are observed to be significant (considering the p-values). This model has been finalised due to its high performance.

Model:

```
Call:
lm(formula = SalePrice ~ LotArea + housefunction + exterior_cond +
    WoodDeckSF + kitchen + PoolArea + OverallCond + bsmt_exp +
    MSSubClass + GarageCars + sale_cond + nbhd_price_level +
    OverallQual + X2ndFlrSF + X1stFlrSF + culdesac_fr3 + MasVnrArea,
    data = training)
```

Residual Analysis:

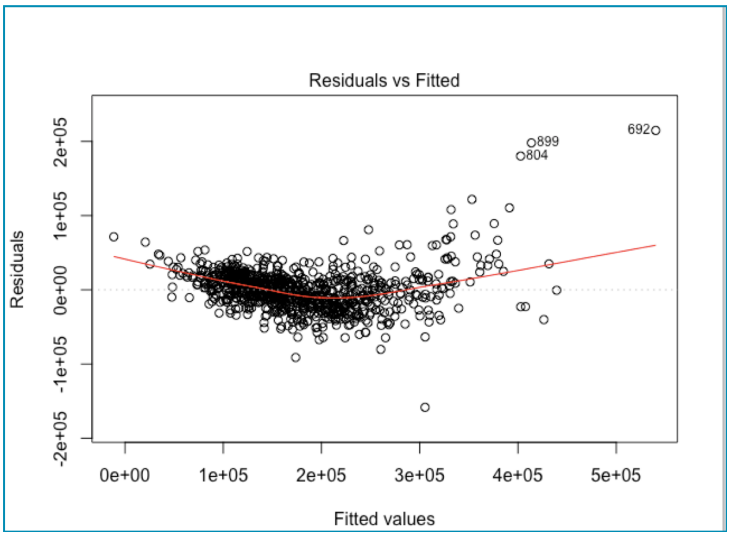


Fig: Residuals (vs) Fitted plot

From the above plot, we can infer that

- The graph is not completely mimicking a Funnel pattern.
- This mean it slightly tends to violate the independence assumption.

The RMSE value for the final model is 10.58682.

Regression Tree:

Introduction: The linear regression does not give visual interpretation feature for prediction of the house prices but the regression tree does. The tree structure is where each node represents a question and the continuation from the node ends to point representing the answers, they are also called edges.

Model:

```
Regression tree:
tree(formula = SalePrice ~ LotArea + housefunction + exterior_cond +
    WoodDeckSF + kitchen + PoolArea + OverallCond + bsmt_exp +
    MSSubClass + GarageCars + sale_cond + nbhd_price_level +
    OverallQual + X2ndFlrSF + X1stFlrSF + culdesac_fr3 + MasVnrArea,
    data = training)
Variables actually used in tree construction:
[1] "OverallQual"      "nbhd_price_level" "X2ndFlrSF"        "X1stFlrSF"
[5] "exterior_cond"    "bsmt_exp"
Number of terminal nodes:  11
```

From the regression tree result, we can find that the model uses only 7 variables out of the 17 selected one. It means that the importance of these 7 variables are highest and contribute the most in predicting the house prices. Let's analyze the plot:

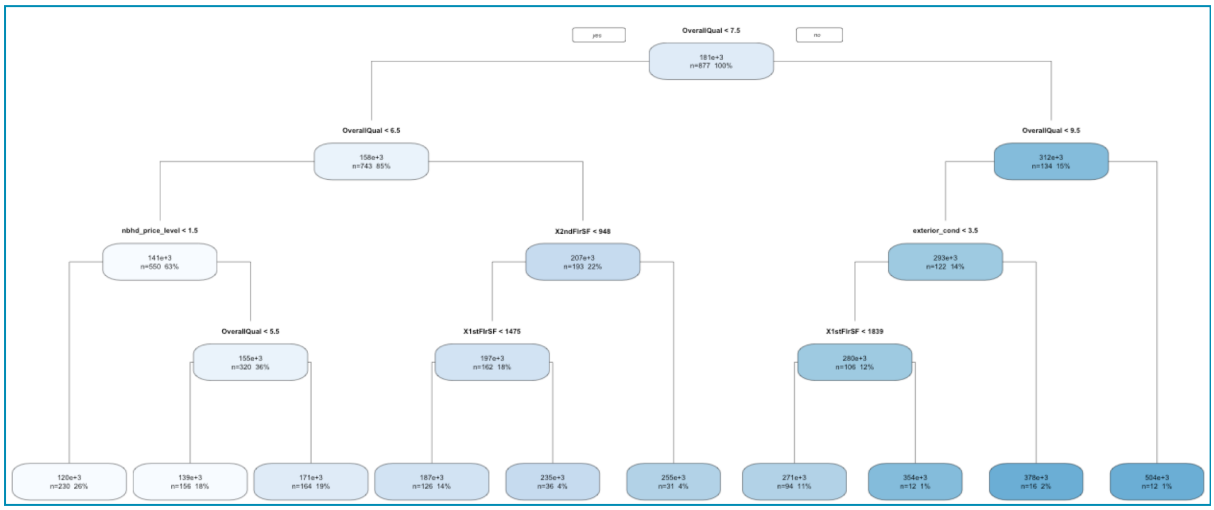


Fig: Regression Tree.

On analysing the tree, If the predictor "over- all condition" is greater than 7.5(the overall quality are Integers) and the "over-all quality" is greater than 9.5 then the house price is predicted to be 504000$. AS you can see that the prediction is limited to the options available on the tree. Although the tree method gives us the advantage of interpretation, it is achieved by compensating it with variance. The output accuracy is lost.

The RMSE of this model is 10.80467.

Random Forest:

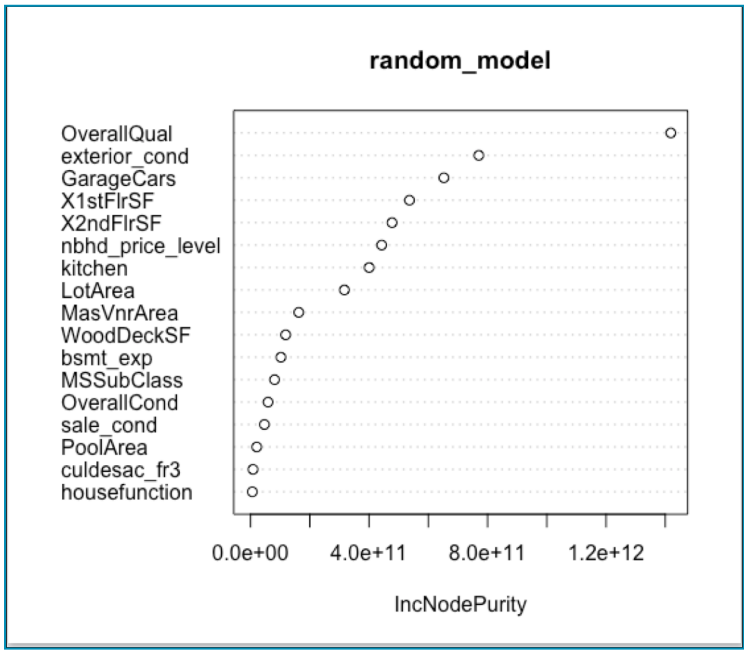
Introduction: The random Forest method improves the accuracy of the tree method. This is done by running continues validation process until statistically significant and appreciable accuracy of the model is determined.

Model:

```
Call:
 randomForest(formula = SalePrice ~ LotArea + housefunction +      exterior_cond + WoodDeckSF + kitchen + PoolArea + OverallCond +      bsmt_exp + MS
SubClass + GarageCars + sale_cond + nbhd_price_level +      OverallQual + X2ndFlrSF + X1stFlrSF + culdesac_fr3 + MasVnrArea,      data = training)
               Type of random forest: regression
                     Number of trees: 500
No. of variables tried at each split: 5

          Mean of squared residuals: 893127481
                    % Var explained: 86.35
```

This is basically collection of tree models and the random is the best out of it. In the Random model we can plot a graph called importance plot:



As per this random forest model, we can see that the most important variable is Overall Quality, that is the selection is of ratings from 1-10. The variables importance plot is different from tree model because it uses multiple models for better performance than just from a single tree model.

The RMSE value of this model is 10.36557.

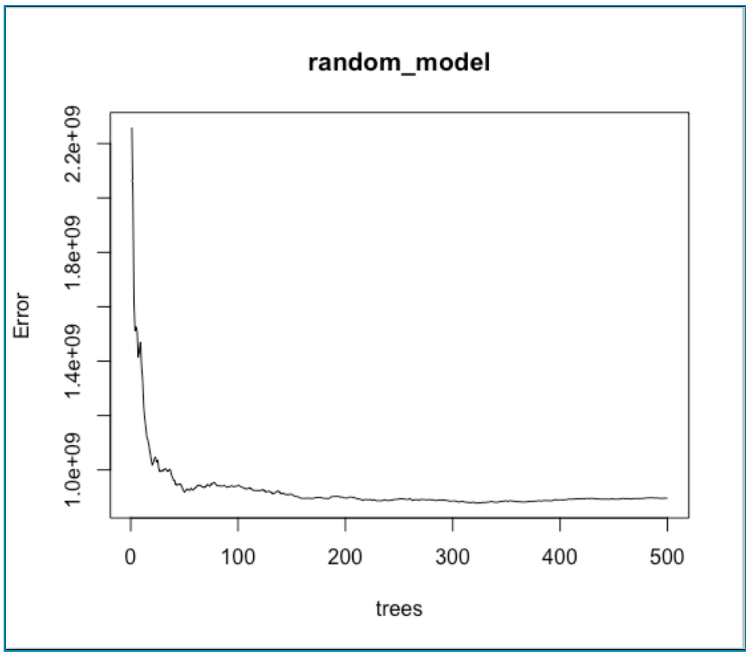


Fig: RMSE (vs) No.of trees.

The above figure is comparison between RMSE values and Number of trees. In this random forest model the error flattens at two hundred trees.

Final Results:

Model	RMSE
Linear Modelling	10.66526
Backward Model	10.66511
Final Model	10.58682
Regression Tree	10.80467
Random Forest Model	10.36557

The Random forest model stands out as the final winner with the least RMSE (10.36).

Conclusion:

To predict the house prices different models were used like the Linear regression, Regression tree and the Random forest. The Random forest model yielded the best performance with an RMSE of 10.36557. It is evident that different models lead to different results. We could also infer that the importance of predictors governing the house price are also the same across the models. The most important ones are the “Overall condition”, “Garage Cars” – number of cars that can be parked, “Area of the second floor”, “Neighbourhood type”, “lot area” and “the basement exposure. Three algorithms were used on the house price prediction data. The Random Forest method produced some good results and accuracy and hence this algorithm can be used to predict the house prices. The models accuracy and performance were compared between models using Root Mean Square Error (RMSE). If large property broking companies adopt such models they can assign or predict the house prices without the intervention of the third-party brokers. These models can save time, money and reduce the dependency of human intervention. Other than the main objective of predicting the house prices, the important variables or factors that impact the price of a house have thus been found out.