DSP - 08
Group- 4

14-07-18

CAPSTONE PROJECT :

Predictive Analytics on Ames Housing Data.

Submitted By:

Chetan Jawale.

ABSTRACT

The goal of this project is to analyze the given Housing data and predict the prices of houses precisely. The data was divided into two parts i.e the training and testing dataset respectively.

Models were developed based on training dataset excluding some observations which were keptfor validation purpose, and then applied on the test dataset after finding accuracy of perticular models to predict the respective values.

Based on the prices predicted, by each individual model we can determine best fitting model as per our objective.

TABLE OF CONTENTS

INTRODUCTION

Ames Housing Authority is a public housing agency that serves the city of Ames, Iowa, US. It helps provide decent and safe rental housing for eligible low-income families, the elderly, and persons with disabilities.The housing authority has collected 79 assessment parameters which describes every aspect of residential homes in Ames. These variables focus on the quality and quantity of the physical attributes of a property. Most of the variables are exactly the type of information that a typical home buyer would want to know about a potential property.

The project uses Feature selection method of Random Forest Selector (RFS) by application of package Boruta. To predict the house prices in city Ames, The machine learning algorithms used are, Linear Regression, Support Vector Machines, Generalised Linear Model and Random Forest Model.
With the help of analytics on this real-world housing data, one can easily get familier with features those are most important in determining housing prices in city of Ames.

Number of features in the given dataset: 80

Number of observations in t he given dataset: 1460

The Dataset :

We have,

1) a training set which contains data about houses and their prices,
2) a test set which contains data about a different set of houses, for which we will predict SalePrice.

< str(data)

```
'data.frame':   2919 obs. of  81 variables:
 $ Id           : int  1 2 3 4 5 6 7 8 9 10 ...
 $ MSSubClass   : int  60 20 60 70 60 50 20 60 50 190 ...
 $ MSZoning     : Factor w/ 5 levels "C (all)","FV",..: 4 4 4 4 4 4 4 4 5 4 ...
 $ LotFrontage  : int  65 80 68 60 84 85 75 NA 51 50 ...
 $ LotArea      : int  8450 9600 11250 9550 14260 14115 10084 10382 6120 7420 ...
 $ Street       : Factor w/ 2 levels "Grvl","Pave": 2 2 2 2 2 2 2 2 2 2 ...
 $ Alley        : Factor w/ 2 levels "Grvl","Pave": NA NA NA NA NA NA NA NA NA NA ...
 $ LotShape     : Factor w/ 4 levels "IR1","IR2","IR3",..: 4 4 1 1 1 1 4 1 4 4 ...
 $ LandContour  : Factor w/ 4 levels "Bnk","HLS","Low",..: 4 4 4 4 4 4 4 4 4 4 ...
 $ Utilities    : Factor w/ 2 levels "AllPub","NoSeWa": 1 1 1 1 1 1 1 1 1 1 ...
 $ LotConfig    : Factor w/ 5 levels "Corner","CulDSac",..: 5 3 5 1 3 5 5 1 5 1 ...
 $ LandSlope    : Factor w/ 3 levels "Gtl","Mod","Sev": 1 1 1 1 1 1 1 1 1 1 ...
 $ Neighborhood : Factor w/ 25 levels "Blmngtn","Blueste",..: 6 25 6 7 14 12 21 17 18 4 ...
 $ Condition1   : Factor w/ 9 levels "Artery","Feedr",..: 3 2 3 3 3 3 3 5 1 1 ...
 $ Condition2   : Factor w/ 8 levels "Artery","Feedr",..: 3 3 3 3 3 3 3 3 3 1 ...
 $ BldgType     : Factor w/ 5 levels "1Fam","2fmCon",..: 1 1 1 1 1 1 1 1 1 2 ...
 $ HouseStyle   : Factor w/ 8 levels "1.5Fin","1.5Unf",..: 6 3 6 6 6 1 3 6 1 2 ...
 $ OverallQual  : int  7 6 7 7 8 5 8 7 7 5 ...
 $ OverallCond  : int  5 8 5 5 5 5 5 6 5 6 ...
 $ YearBuilt    : int  2003 1976 2001 1915 2000 1993 2004 1973 1931 1939 ...
 $ YearRemodAdd : int  2003 1976 2002 1970 2000 1995 2005 1973 1950 1950 ...
 $ RoofStyle    : Factor w/ 6 levels "Flat","Gable",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ RoofMatl     : Factor w/ 8 levels "ClyTile","CompShg",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ Exterior1st  : Factor w/ 15 levels "AsbShng","AsphShn",..: 13 9 13 14 13 13 13 7 4 9 ...
 $ Exterior2nd  : Factor w/ 16 levels "AsbShng","AsphShn",..: 14 9 14 16 14 14 14 7 16 9 ...
 $ MasVnrType   : Factor w/ 4 levels "BrkCmn","BrkFace",..: 2 3 2 3 2 3 4 4 3 3 ...
 $ MasVnrArea   : int  196 0 162 0 350 0 186 240 0 0 ...
 $ ExterQual    : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 4 3 4 3 4 4 4 ...
 $ ExterCond    : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ Foundation   : Factor w/ 6 levels "BrkTil","CBlock",..: 3 2 3 1 3 6 3 2 1 1 ...
 $ BsmtQual     : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 3 3 4 3 3 1 3 4 4 ...
 $ BsmtCond     : Factor w/ 4 levels "Fa","Gd","Po",..: 4 4 4 2 4 4 4 4 4 4 ...
 $ BsmtExposure : Factor w/ 4 levels "Av","Gd","Mn",..: 4 2 3 4 1 4 1 3 4 4 ...
 $ BsmtFinType1 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 3 1 3 1 3 3 3 1 6 3 ...
 $ BsmtFinSF1   : int  706 978 486 216 655 732 1369 859 0 851 ...
 $ BsmtFinType2 : Factor w/ 6 levels "ALQ","BLQ","GLQ",..: 6 6 6 6 6 6 6 2 6 6 ...
 $ BsmtFinSF2   : int  0 0 0 0 0 0 0 32 0 0 ...
 $ BsmtUnfSF    : int  150 284 434 540 490 64 317 216 952 140 ...
 $ TotalBsmtSF  : int  856 1262 920 756 1145 796 1686 1107 952 991 ...
 $ Heating      : Factor w/ 6 levels "Floor","GasA",..: 2 2 2 2 2 2 2 2 2 2 ...
 $ HeatingQC    : Factor w/ 5 levels "Ex","Fa","Gd",..: 1 1 1 3 1 1 1 1 3 1 ...
 $ CentralAir   : Factor w/ 2 levels "N","Y": 2 2 2 2 2 2 2 2 2 2 ...
 $ Electrical   : Factor w/ 5 levels "FuseA","FuseF",..: 5 5 5 5 5 5 5 5 2 5 ...
 $ X1stFlrSF    : int  856 1262 920 961 1145 796 1694 1107 1022 1077 ...
 $ X2ndFlrSF    : int  854 0 866 756 1053 566 0 983 752 0 ...
 $ LowQualFinSF : int  0 0 0 0 0 0 0 0 0 0 ...
 $ GrLivArea    : int  1710 1262 1786 1717 2198 1362 1694 2090 1774 1077 ...
 $ BsmtFullBath : int  1 0 1 1 1 1 1 1 0 1 ...
 $ BsmtHalfBath : int  0 1 0 0 0 0 0 0 0 0 ...
 $ FullBath     : int  2 2 2 1 2 1 2 2 2 1 ...
 $ HalfBath     : int  1 0 1 0 1 1 0 1 0 0 ...
 $ BedroomAbvGr : int  3 3 3 3 4 1 3 3 2 2 ...
 $ KitchenAbvGr : int  1 1 1 1 1 1 1 1 2 2 ...
 $ KitchenQual  : Factor w/ 4 levels "Ex","Fa","Gd",..: 3 4 3 3 3 4 3 4 4 4 ...
 $ TotRmsAbvGrd : int  8 6 6 7 9 5 7 7 8 5 ...
 $ Functional   : Factor w/ 7 levels "Maj1","Maj2",..: 7 7 7 7 7 7 7 7 3 7 ...
 $ Fireplaces   : int  0 1 1 1 1 0 1 2 2 2 ...
 $ FireplaceQu  : Factor w/ 5 levels "Ex","Fa","Gd",..: NA 5 5 3 5 NA 3 5 5 5 ...
 $ GarageType   : Factor w/ 6 levels "2Types","Attchd",..: 2 2 2 6 2 2 2 2 6 2 ...
 $ GarageYrBlt  : int  2003 1976 2001 1998 2000 1993 2004 1973 1931 1939 ...
 $ GarageFinish : Factor w/ 3 levels "Fin","RFn","Unf": 2 2 2 3 2 3 2 2 3 2 ...
 $ GarageCars   : int  2 2 2 3 3 2 2 2 2 1 ...
 $ GarageArea   : int  548 460 608 642 836 480 636 484 468 205 ...
 $ GarageQual   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 2 3 ...
 $ GarageCond   : Factor w/ 5 levels "Ex","Fa","Gd",..: 5 5 5 5 5 5 5 5 5 5 ...
 $ PavedDrive   : Factor w/ 3 levels "N","P","Y": 3 3 3 3 3 3 3 3 3 3 ...
 $ WoodDeckSF   : int  0 298 0 0 192 40 255 235 90 0 ...
 $ OpenPorchSF  : int  61 0 42 35 84 30 57 204 0 4 ...
 $ EnclosedPorch: int  0 0 0 272 0 0 0 228 205 0 ...
 $ X3SsnPorch   : int  0 0 0 0 0 320 0 0 0 0 ...
 $ ScreenPorch  : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PoolArea     : int  0 0 0 0 0 0 0 0 0 0 ...
 $ PoolQC       : Factor w/ 3 levels "Ex","Fa","Gd": NA NA NA NA NA NA NA NA NA NA ...
 $ Fence        : Factor w/ 4 levels "GdPrv","GdWo",..: NA NA NA NA NA 3 NA NA NA NA ...
 $ MiscFeature  : Factor w/ 4 levels "Gar2","Othr",..: NA NA NA NA NA 3 NA 3 NA NA ...
 $ MiscVal      : int  0 0 0 0 0 700 0 350 0 0 ...
 $ MoSold       : int  2 5 9 2 12 10 8 11 4 1 ...
 $ YrSold       : int  2008 2007 2008 2006 2008 2009 2007 2009 2008 2008 ...
 $ SaleType     : Factor w/ 9 levels "COD","Con","ConLD",..: 9 9 9 9 9 9 9 9 9 9 ...
 $ SaleCondition: Factor w/ 6 levels "Abnorml","AdjLand",..: 5 5 5 1 5 5 5 5 1 5 ...
 $ SalePrice    : num  208500 181500 223500 140000 250000 ...
```

Summary of the data/ Review of Literature :

The data types of the columns are mixed: we have integers, numric data & factors (levels). So, it's clear that features come in fundamentally different types :

1. Some features are inherently **NUMERICAL**. They are quantities that we can measure or count. Some of these are continuous, such as the total living area (GrLivArea), while others are discrete, such as the number of rooms (TotRmsAbvGrd).

2. Other features are **CATEGORICAL**. They are qualitative or descriptive in nature. For example, this includes the neighbourhood in which the house is located (Neighborhood), and the type of foundation the house was built on (Foundation). There is no inherent ordering to these features.

3. Yet others are **ORDINAL**. They comprise categories with an implicit order. Examples of this include the overall quality rating (OverallQual) or the irregularity of the lot (LotShape). We can think of them as representing values on an arbitrary scale.

In total, the training set contains 1460 rows, each of these represents one house sold. Some columns, however, contain notably fewer entries. This tells us that we have **MISSING VALUES** in our dataset. But, some of the features represents NA's as if 'that perticular feature is not available', leads us to re-level such features. These features are:

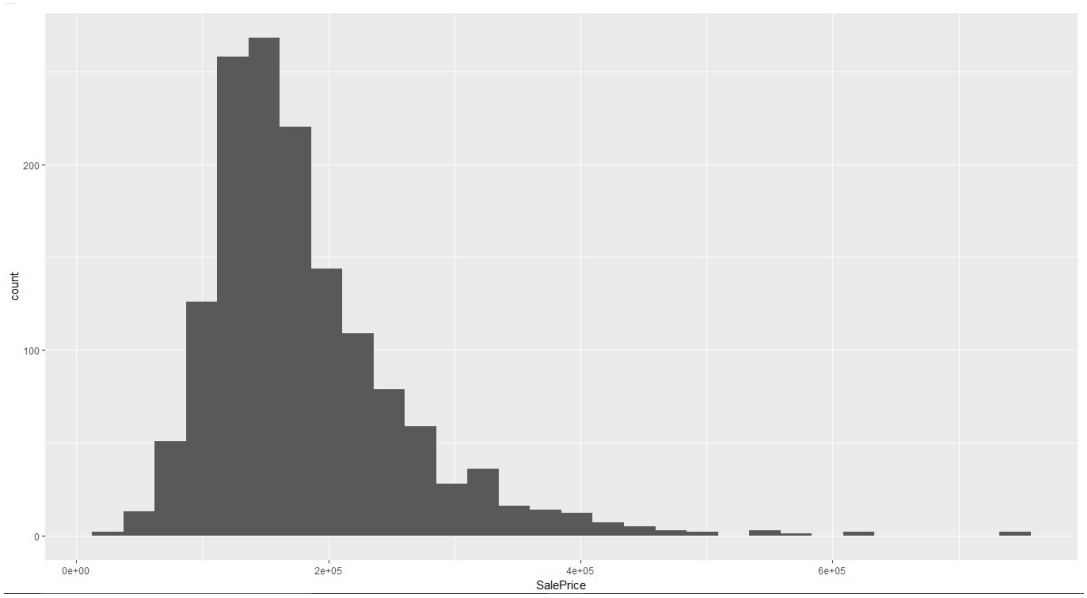
("Alley","BsmtQual","BsmtCond","BsmtExposure","BsmtFinType1","BsmtFinType2","FireplaceQu","GarageType","GarageFinish","GarageQual","GarageCond","PoolQC","Fence","MiscFeature")

Rest of the missing values were imputed by the Means & Modes in the numeric data & the categorical data respectively by using **For-Loop** to fasten the process and to ensure accurate analysis.

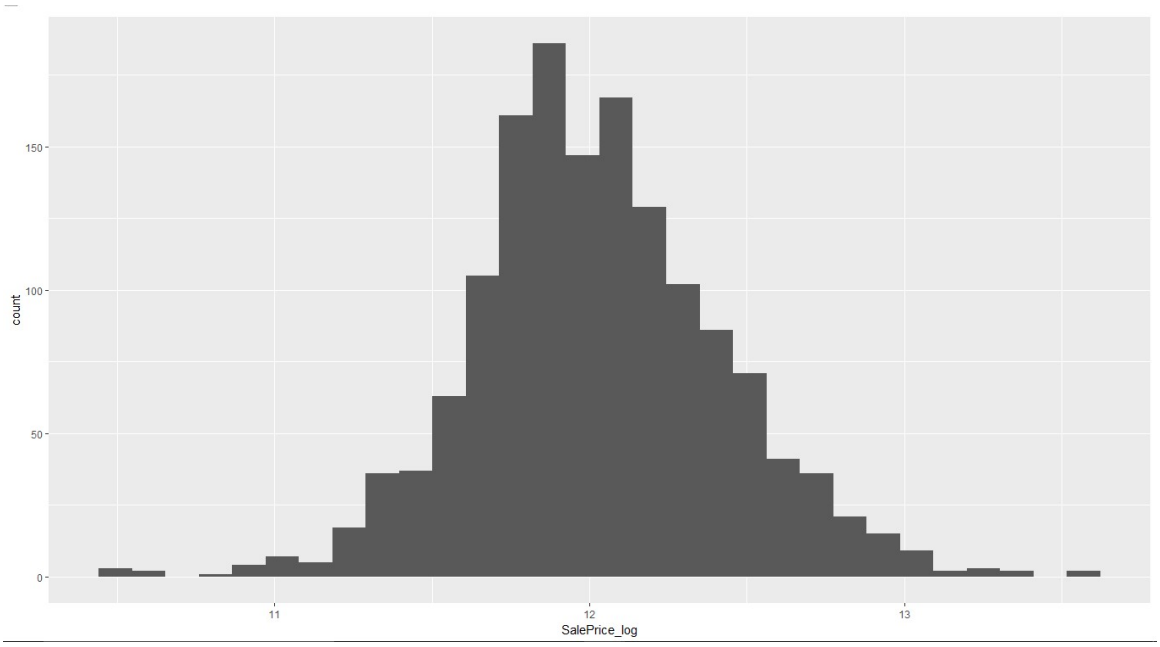
DATA VISUALIZATION :

Univariate Analysis on Numerical Variables :

Histograms are perfect for visualizing distribution of numeric variables. Below is the distribution plotted on target variable 'SalePrice'.

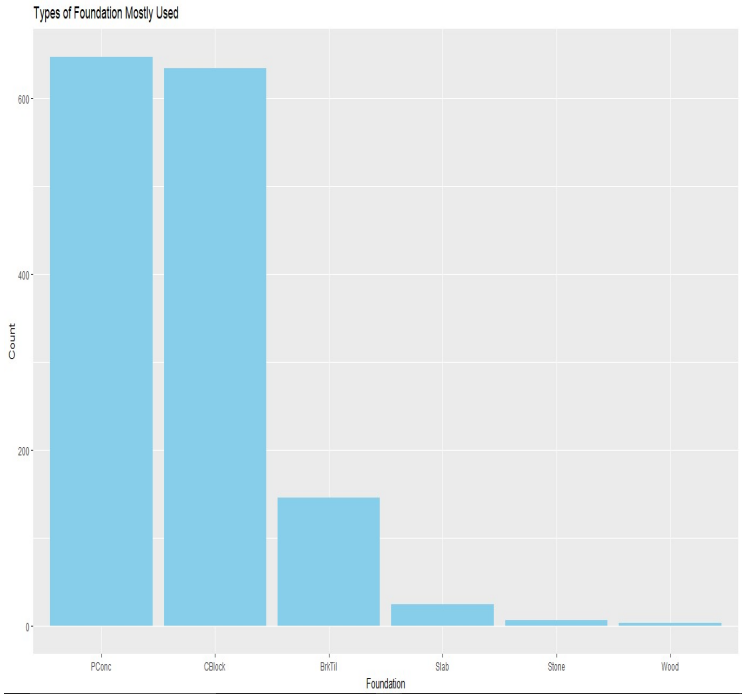


We notice that the distribution is skewed towards cheaper homes, with a relatively long tail at high prices. To make the distribution more symmetric, we can try taking its logarithm.



Besides making the distribution more symmetric, working with the log of the sale price will also ensure that relative errors for cheaper and more expensive homes are treated on an equal footing. As such, we can think of taking 'log(SalePrice)' as our true target variable.

Univariate Analysis on Categorical Variables :

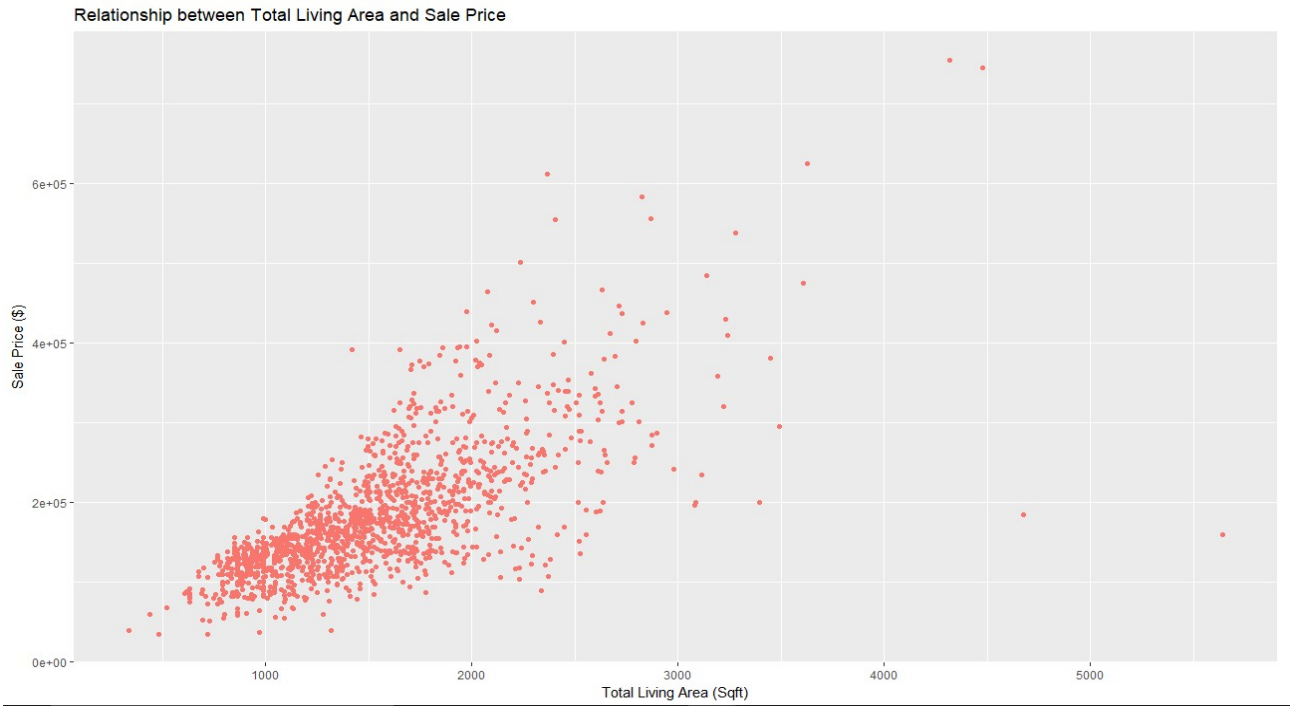


Here is a bar chart representing an important feature **'Foundation'** which contains types of materials used in construction such as, 'Brick & Tile', 'Cinder Block', 'Poured Contrete', 'Slab', 'Stone, Wood'.

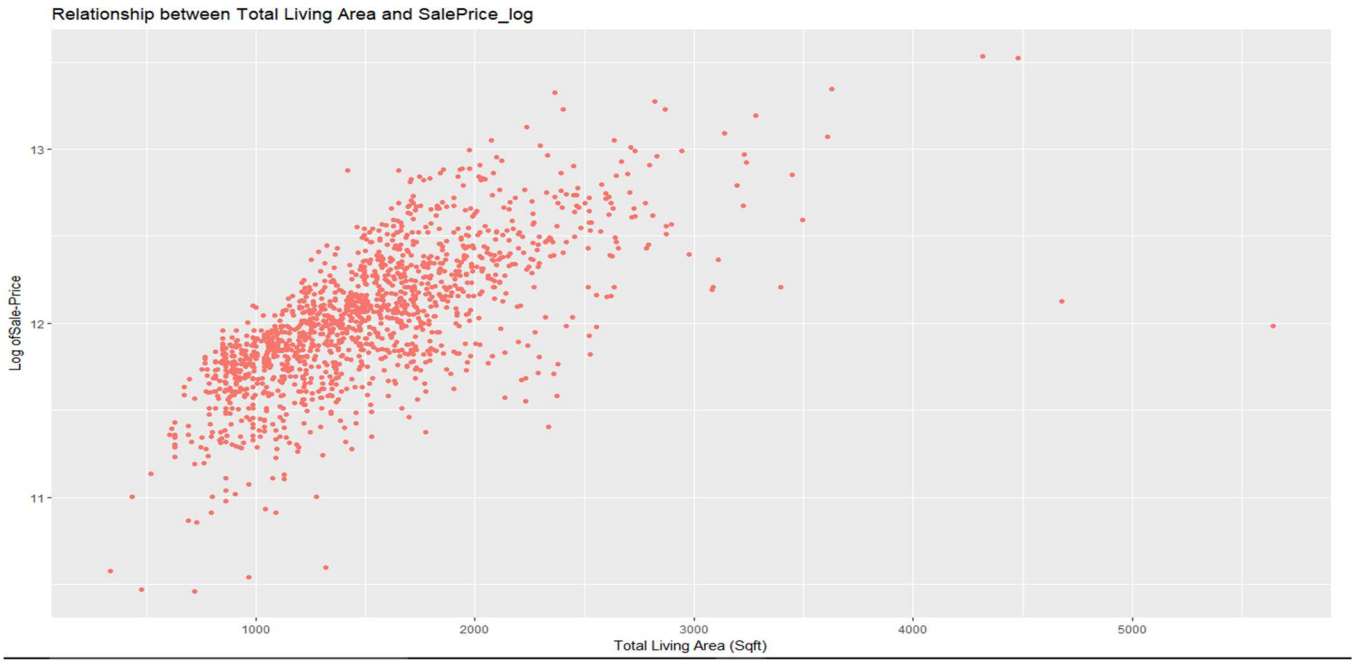
From the bar chart we can say, Stone and wood are very rarely used materials. Wheras Poured Concrete & Cinder Blocks are extensively used.

Bivariate Analysis on Numerical Variables :

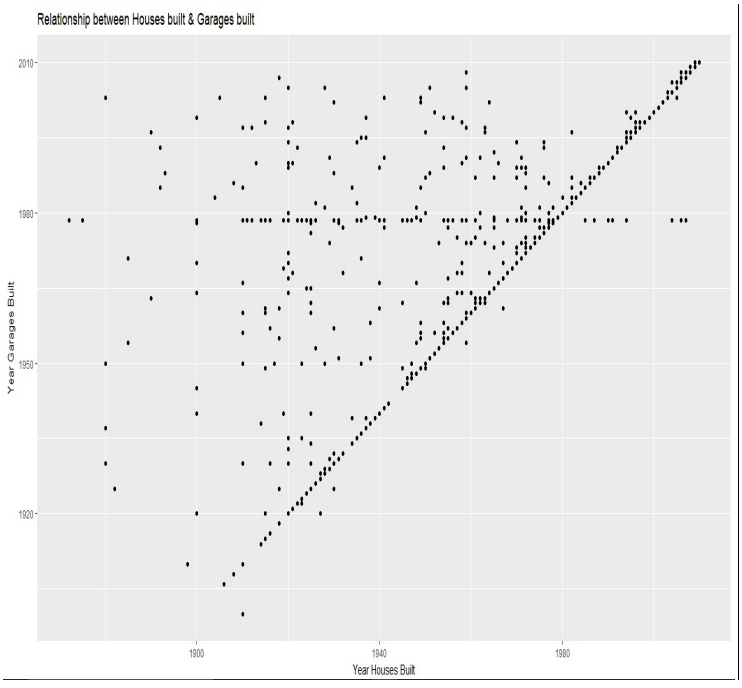
Since, The total living area of a house is likely to be an important factor in determining its price. Here is a Scatter plot explaining the relationship between 'GrLivArea' and 'SalePrice'.



There is indeed a strong dependence of ‘Saleprice’ on the ‘Total living area’,As expected. The larger the house, The more expensive it tends to be.
When we take the log in the second plot, the distribution looks notably more balanced.

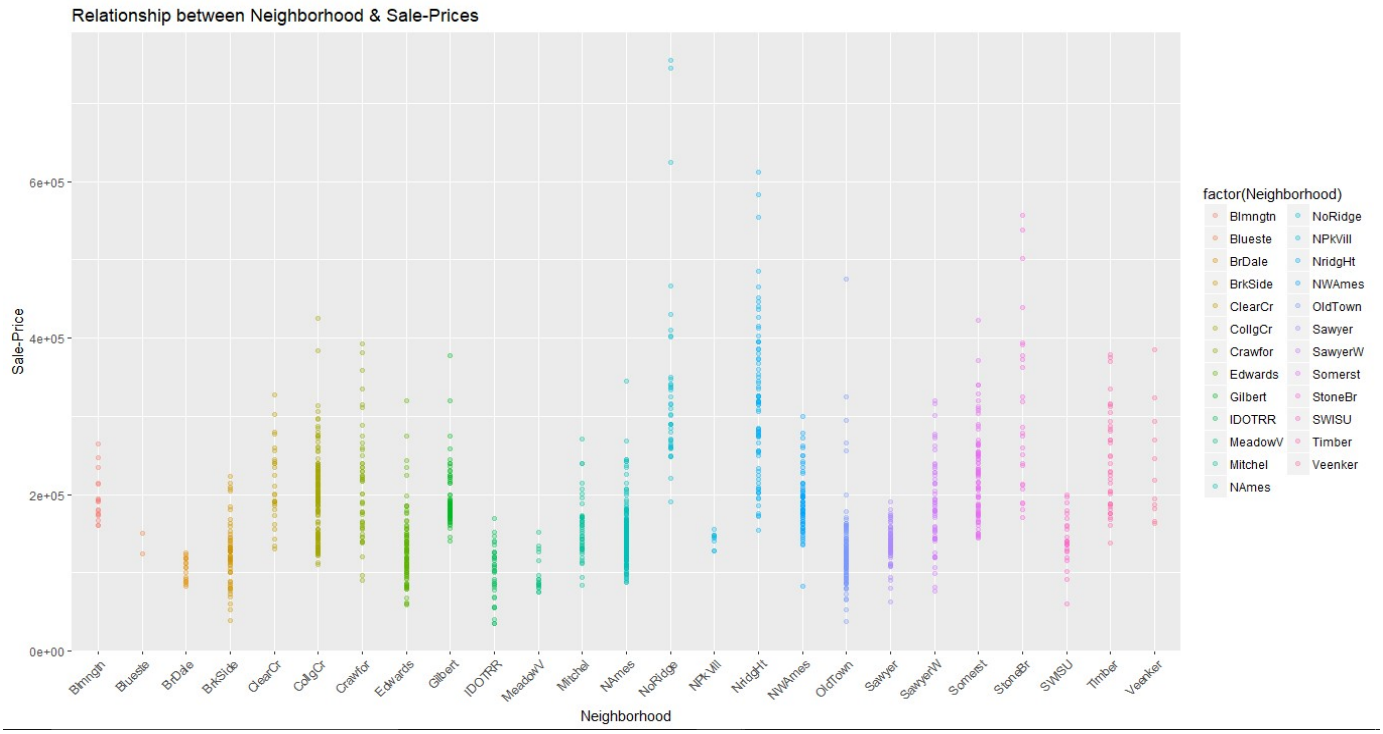


There are two points that don’t seem to fit in with the rest. Towards the lower right part of the plot, there are two very large houses (bigger than 4500 sqft) with unusually low sale prices**.** Hence, we conclude there are **OUTLIERS** present in the dataset.



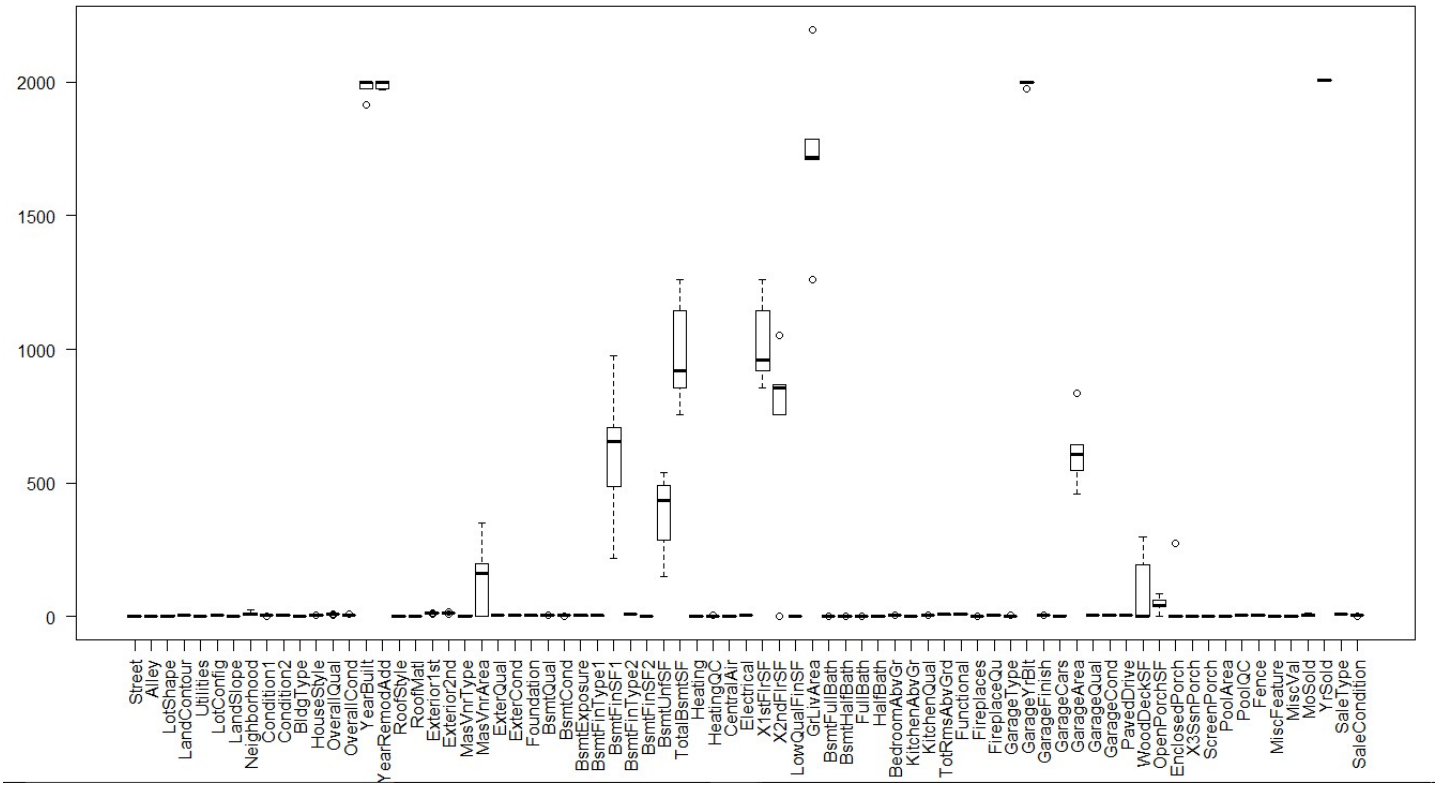
Scatter plot of ‘YearBuilt’ and ‘GarageYrBlt’ tells us that the majority of garages were built at the same time as the houses they belong to; these forms the diagonal line that runs across the plot. A significant number were also added later, these are the points above the line.

Bivariate Analysis on Categorical Variables :



From a graph above we can say, there is considerable variation in price between neighbourhoods. The figure allows to get an idea of how different areas compare to each other at a glance.

Boxplot of Data :



Feature Selection :

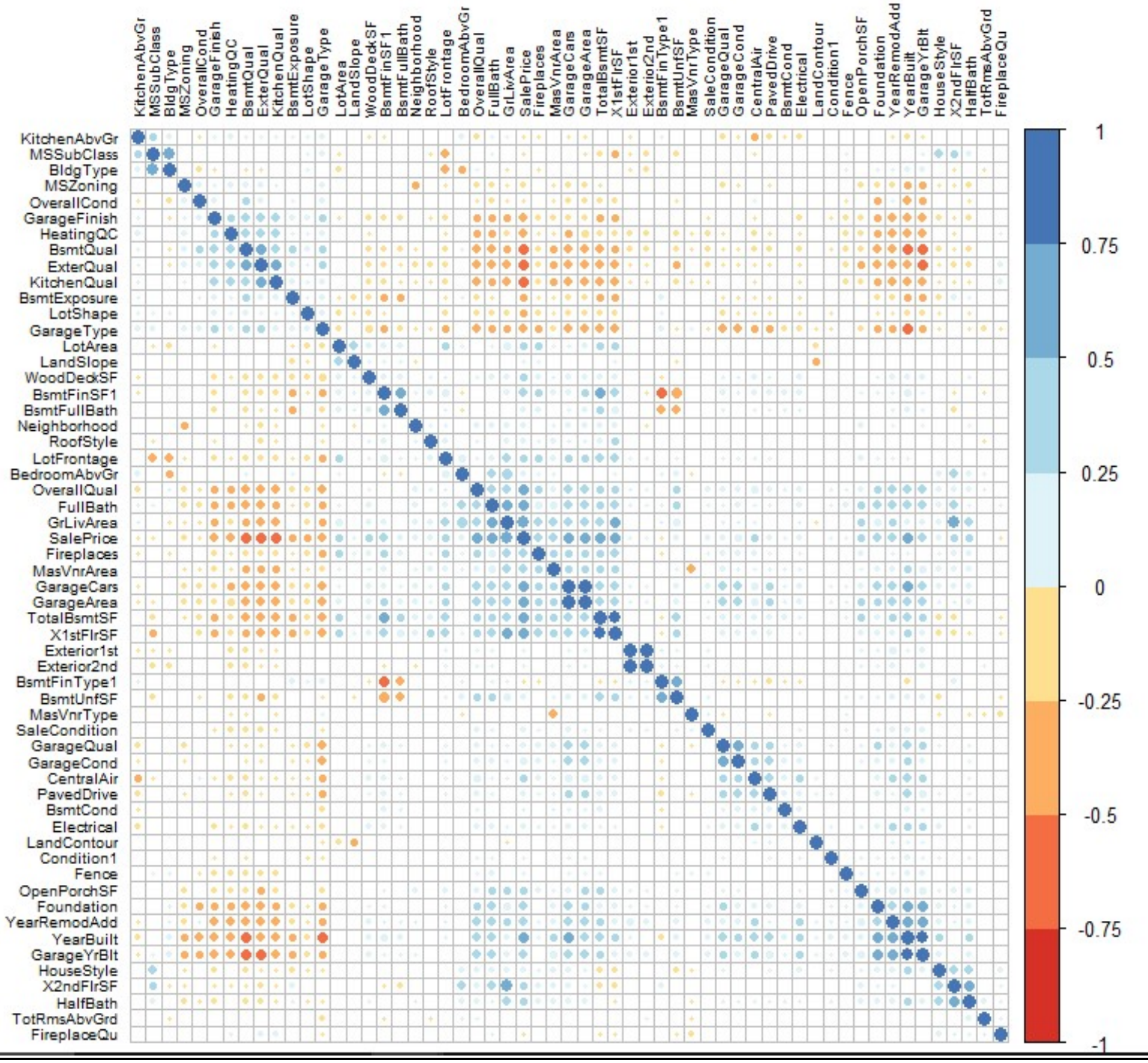
Feature selection is considered one of the important steps in the predictive modelling.

Boruta is a wrapper algorithm for all relevant feature selection, capable of working with any classi-fication method that output variable importance measure (VIM); by default, Boruta uses RandomForest. The method performs a top-down search for relevant features by comparing original attributes' importance with importance achievable at random, estimated using their permuted copies, and progressively elliminating irrelevant featurs to stabilise that test.

Features	normHits	Decision
MSSubClass	1.00000000	Confirmed
MSZoning	1.00000000	Confirmed
LotFrontage	1.00000000	Confirmed
LotArea	1.00000000	Confirmed
Street	0.00000000	Rejected
Alley	0.48484848	Rejected
LotShape	0.69696970	Confirmed
.	.	.
.	.	.
.	.	.
.	.	.
MoSold	0.00000000	Rejected
YrSold	0.01010101	Rejected
SaleType	0.01010101	Rejected
SaleCondition	0.63636364	Confirmed

Output showing feature importance

The Following figure shows the CORRELATION MATRIX HEATMAP in r using package corrplot. This gives us an idea about which features are highly correlated and are considerable or not.



STEPS PERFORMED :

1. Loading the data into the Rstudio.

```
ames_train <- read.csv(file.choose())
ames_test <- read.csv(file.choose())
```

2. Making the necessary conversions to make the data more usable.
   a. Introduced level "NO" to features where NA's are not supposed to consider missing values.
   b. Log transformed the target feature in order to achieve normalcy.
   c. Converted ordinal variables such as 'OverallQual','OverallCond','TotRmsAbvGrd' into the categorical variables.
   d. Took a Subset of important features selected by algorithm to build a model on.

3. Cleaning data.
   a. Treated missing values by imputing Means and Modes in Numeric variables and Categorical variables respectively.

4. Visualized data, by developing various plots for exploratory analysis.

5. Creating a Training and Validation dataset with 75% being the training data and 25% being the validation data.

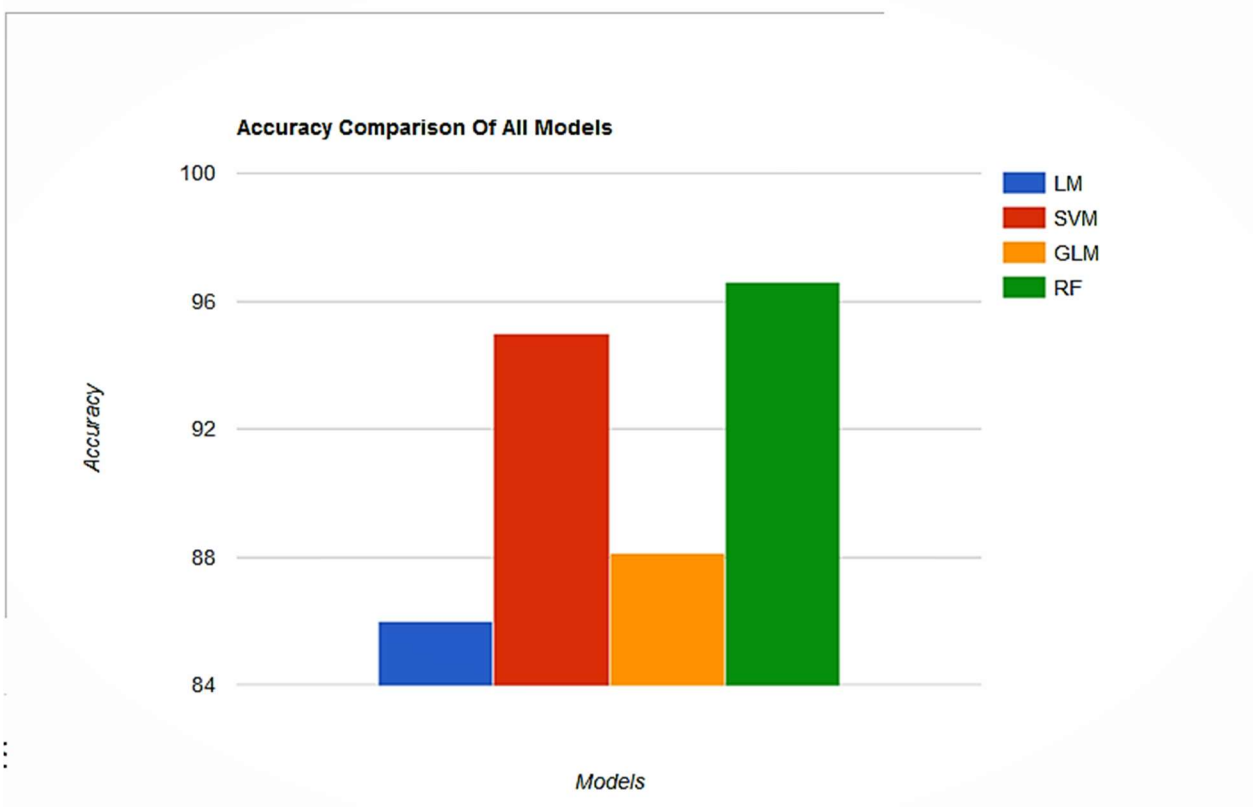
6. Building Models.
   a. **Model 1 :** Linear regression algorithm. It establishes linear relationship between the predictor variables(x) and response variable(y) to predict for the unknown values in response variable when the predictor values are known.
   b. **Model 2 :** Support Vector Machines algorithm. It is a data classification method that separates data using hyperplanes.
   c. **Model 3 :** Generalised Linear Model. It is used to fit generalised models, specified by giving a symbolic description of the linear predictor and a desription of the error distribution.
   d. **Model 4 :** RandomForest algorithm. It is a tree-based algorithm which involves building several trees (decision trees), then combining their output to improve generalization ability of the model.

7. Validating builded models on the validation set and creating the actual confusion matrix based on predicted values and analyzing the results.

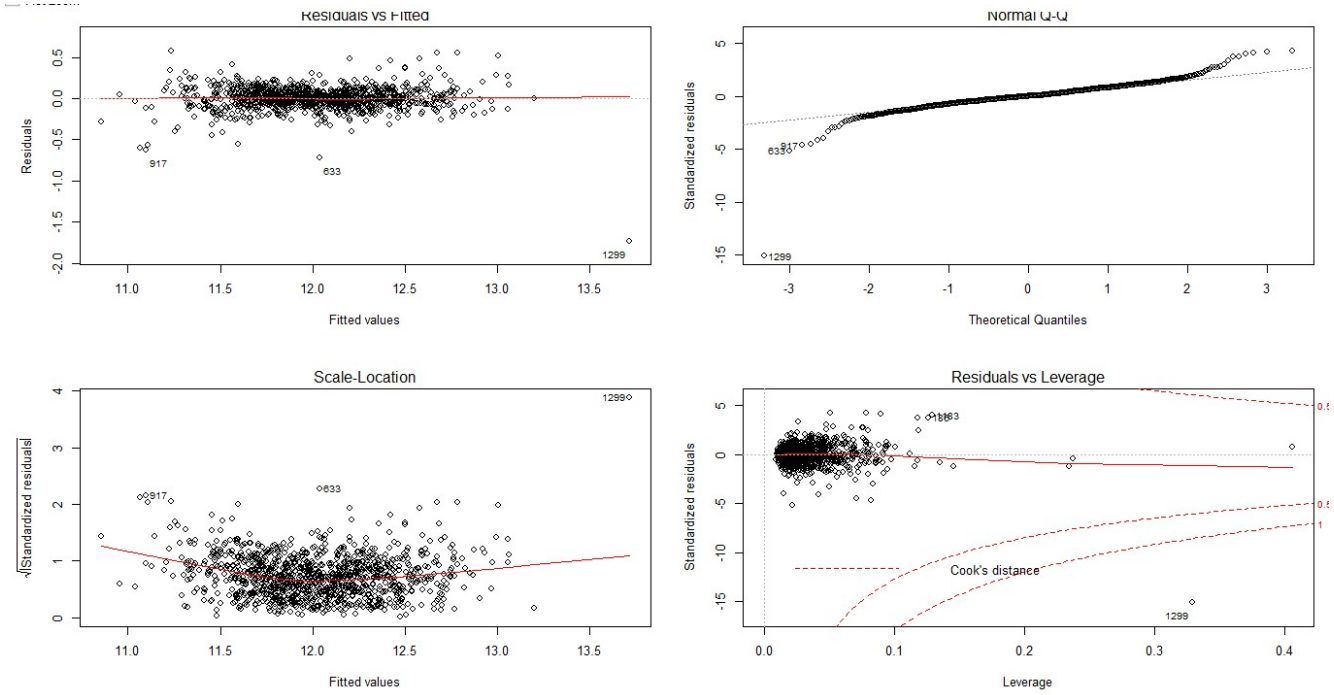
8. Predicting values for the testing dataset and Interpreting results.

Model Evaluation :

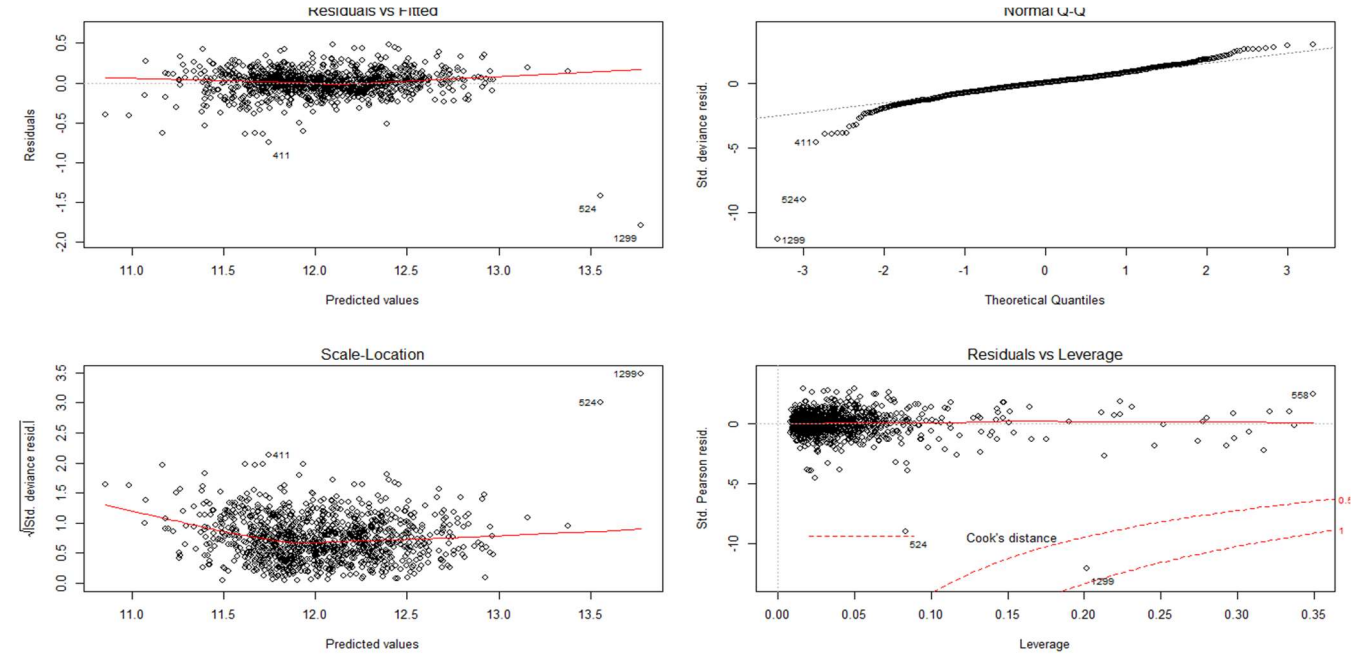
Model	Median Prediction Error in '$'	% Difference than Correct Value
Random Forest	39599.94	2.954111
Support Vector Machines	55607.61	5.442585
LinearModel	65514.39	7.91715
Generalised Linear Model	66106.05	8.290854



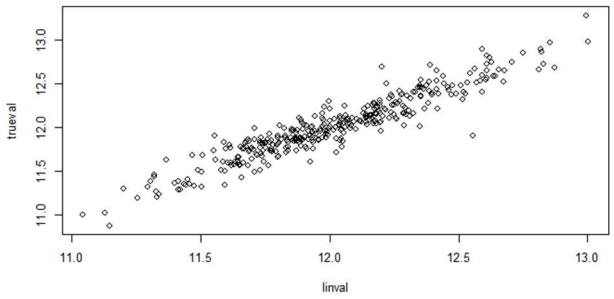
Model Comparison (LM vs GLM) :



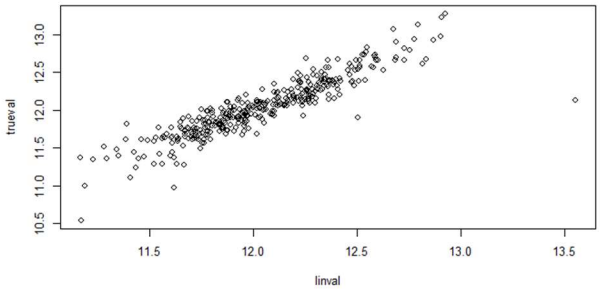
Residual VS Fitted and Normality (Linear Model)



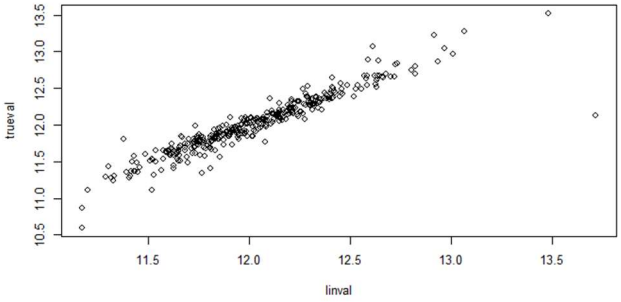
Residual VS Fitted and Normality (Generalised Linear Model)



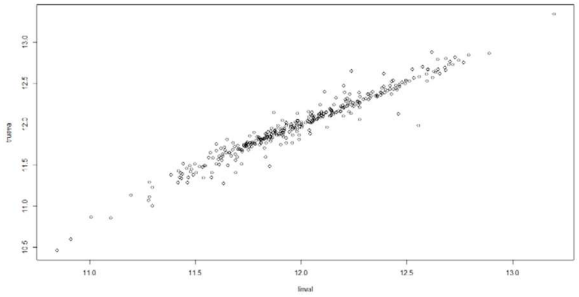
1 GLM Model



2. LM Model



3 SVM Model



4. RF Model

Correlation Plot of Actuals vs Predicted Prices of all models

As we can see that the Random Forest Model (Model 4) is highly accurate in predicting precise house prices. i.e 97 % and seems to be the best of all among models we have trained.

(Model 3) SVM is next best in predicting house prices after RFM with 94% accuracy.

However, Generalised Linear Model and Linear Model scores nearby same accuracy between 90% to 92%.

The Prediction accuracy can be further improved by using ensemble modelling.

Conclusion :

The relationship between house prices and the economy is an important factor for predicting house prices. As per buyer and sellers concern Housing prices trends are very important to study before making an investment, Hence it is directly or indirectly related to current economic situation. Therefore it is important to predict housing prices without bias to help both buyers and sellers make their decisions.

The data contains list of helpful features as well as the unneccesary or luxurious features of which probabilities of occurence are very less. The data is more concentrated or helpful for the middle class people or a broker-seller. since, huge amount of data belongs to cheaper or affordable house prices with critial or important aminities.

With the help of Data science and R we have managed to develop a machine learning algorithm to predict housing prices using given features with fair accuracy of around 97% and can be further improved with different possible approaches and ensembeled modelling.

REFERNCES :

1. www.cran.r-project.org (Information on R packages)
2. www.stackoverflow.com (Information on Code optimisation)
3. www.edx.org (Information about visualizing data)
4. www.rapidtables.com (Graphical representation)