After reading the data of Ames Housing, we find that there are 82 attributes and 2930 observations. When we look at the summary of the data, we find many characteristic variables. So, we factor all the attributes with the following command.

midData=as.data.frame(unclass(midData))

Now, by looking at the summary of midData, we find order and PID attributes to be useless. These attributes are simply the serial numbers for the observations. So, we delete the relevant data as it does not add any value to the modelling.

midData$Order=NULL

midData$PID=NULL

Misc.Feature has 2824 NA’s in it. This attribute adds no value as there is no information in this attribute. We delete this attribute as well. Similarly, Fence has 2358 NA’s, Alley has 2732 NA’s, Pool.QC has 2917 NA’s and FirePlace.QU has 1422 NA’s. So, we delete these attributes as well.

Sale.Condition has 2413 observations as Normal. Majority of the data has same value, this will not fetch any good information. So, we delete this attribute. Similarly, Sale.Type has 2536 WD’s, Paves.Drive has 2652 Y’s, Garage.Cond has 2665 TA’s, Garage.Qual has 2615 TA’s, Functional has 2728 typ’s, Electrical has 2682 Sbrkr’s, Central.Air has 2734 Y’s, Heating has 2885 GasA’s, BsmtFin.Type.2 has 2499 Unf’s, Bsmt.Cond has 2616 TA’s, Exter.Cond has 2549 TA’s, Roof.Matl has 2887 CompShg’s, Roof.Style has 2321 Gable’s, Bldg.Type has 2425 1Fam’s, Condition.2 has 2900 Norm’s, Condition.1 has 2522 Norm’s, Land.Slope has 2789 Gtl’s, Utilities has 2927 AllPub’s, Land.Contour has 2633 Lvl’s, Street has 2918 Pave’s, Lot. Config has 2140 Inside’s, MS.Zoning has 2273 RL’s. We delete all these attributes.

Misc.Val has minimum value 0, maximum value as 17000. It’s median is 0 and mean as 50.63. This signifies that the majority of the value is 0 or near to 0. We can easily delete it. Similarly, Pool.Area has maximum 800, minimum and median as 0 and mean as 2.243. Screen.Porch has maximum 576, minimum and median as 0 and mean as 16. X3Ssn.Porch has maximum 508, minimum and median as 0 and mean as 2.592. Enclosed.Porch has maximum 1012, minimum and median as 0 and mean as 23.01.Wood.Deck.SF has maximum1424, minimum and median as 0 and mean as 93.75. Bsmt.Half.Bath has maximum 2, minimum and median as 0 and mean as 0.06113. Low.Qual.Fin.SF has maximum 1064, minimum and median as 0 and mean as 4.677. X2nd.Flr has maximum 2065, minimum and median as 0 and mean as 335.5.BsmtFin.SF.2 has maximum 1526, minimum and median as 0 and mean as 49.72. Mas.Vnr.Area has maximum 1600, minimum and median as 0 and mean as 101.9. We delete all these attributes, as they don’t add any information to the data.

midData$Misc.Feature=NULL

midData$Fence=NULL

midData$Pool.QC=NULL

midData$Fireplace.Qu=NULL

midData$Sale.Condition=NULL

midData$Sale.Type=NULL

midData$Misc.Val=NULL

midData$Pool.Area=NULL

midData$Screen.Porch=NULL

midData$X3Ssn.Porch=NULL

midData$Enclosed.Porch=NULL

midData$Wood.Deck.SF=NULL

midData$Paved.Drive=NULL

midData$Garage.Cond=NULL

midData$Garage.Qual=NULL

midData$Functional=NULL

midData$Bsmt.Half.Bath=NULL

midData$Low.Qual.Fin.SF=NULL

midData$X2nd.Flr.SF=NULL

midData$Electrical=NULL

midData$Central.Air=NULL

midData$Heating=NULL

midData$BsmtFin.SF.2=NULL

midData$BsmtFin.Type.2=NULL

midData$Bsmt.Cond=NULL

midData$Exter.Cond=NULL

midData$Mas.Vnr.Area=NULL

midData$Roof.Matl=NULL

midData$Roof.Style=NULL

midData$Bldg.Type=NULL

midData$Condition.2=NULL

midData$Condition.1=NULL

midData$Land.Slope=NULL

midData$Utilities=NULL

midData$Land.Contour=NULL

midData$Alley=NULL

midData$Street=NULL

midData$Lot.Config=NULL

midData$MS.Zoning=NULL

Now we factorize few numeric data variables as the range of these variables is really low, We check If there are any repetitions for a particular value in the observations for the attributes. We factorize the following attributes: Overall.Cond, Overall.Qual, Bsmt.Full.Bath, Full.Bath, Half.Bath, Bedroom.AbvGr, Kitchen.AbvGr, FirePlaces, Garage.Cars and Yr.Sold. We find that Kitchen.AbvGr has 2796 1’s. So, we delete this variable as well.

midData$Kitchen.AbvGr=NULL

Now the data is ready for modelling. We create models on this processed data.

We model the data using gam.

We get testing Mean Square Error as 324066243.

The mean of SalePrice in Testing Data is 181671.6

The error percent is 9.909 which is quite good. This test gives us a pretty good result with the Generalized Additive Modelling.

In Linear Regression Model using lr, as it cannot predict for factors in the attributes, we rplace with the original data using the following code.

for(j in 1: ncol(midData)){

if(class(midData[,j])=="factor")

midData[,j]=AmesHousing[j]

}

We are getting a testing error of 29420, which is 15.779 percentage.

And we are achieving a testing error of 17.893 percentage.

The adjusted R -squared value is 0.8752

This model is also good with a little higher error rate compared to gam model.

Using glm, we see that the mean square error is 1703689616.

The testing error is 21.9 percentage.

This error is higher than expected. So, it is not optimal to consider this set for modelling this data.