MIDTERM PROJECT

Sowmya Narra (sxn153730)
Tapasya Gutta (txg150730)
Mahesh Paramati (mxp150830)
Manasa Rao Kondrolla (mxk150930)

Aim: To build a model that will predict the Sale Price for the house.

Dataset: There are 81 predictors and the variable to be predicted is “SalePrice”.

- **Number of Instances :** 2930
- **Number of attributes :** 82
- **Number of attributes used for analysis :** 39

APPROACH :

1) **Data Cleaning:** There are many attributes in the dataset which are not predictive which were omitted.

A list of attributes used to build the model are:

MS.Subclass	**MS.Zoning**	**Lot.Frontage**	**Lot.Shape**
Lot.config	**Neighborhood**	**House.style**	**Overall.Qual**
OverAll.Cond	**Year.Built**	**Year.Remod.Add**	**Exterior.1st**
Exterior.2nd	**Exterior.Qual**	**Foundation**	**Bsmt.Exposure**
Bsmt.Qual	**BsmtFin.Type.1**	**BsmtFin.SF.1**	**Bsmr.Unf.SF**
Toatl.Bsmt.Sf	**Heating.QC**	**X1st.Flr.SF**	**Gr.Liv.Area**
Bsmt.Full.Bath	**Bsmt.Half.Bath**	**Full.Bath**	**Half.Bath**
Bedroom.AbvGr	**Kitchen.AbvGr**	**Kitchen.Qual**	**TotRms.AbvGrd**
Garage.Type	**Garage.Yr.Blt**	**Garage.Finish**	**Garage.Cars**
Garage.Area	**Yr.sold**	**Sale Price**	

Table 1 : Attributes considered for Analysis

The following attributes are **not predictive** and hence are **removed**:

1) **Order**: It is unique for each instance and is not required.
2) **PID**: It is unique for each instance and is not required.
3) **Lot Area**: Min. is 1300 and max. is 215245 with a median 9436.
4) **Street**: Grvl is 12 whereas Pave is 2918.
5) **Alley**: There are 2732 NA's.
6) **Land Contour**: One value is 2633 which is very high than others.
7) **Utilities**: One value is 2927 which is very high than others.
8) **Land Slope**: One value is 2789 which is very high than others.
9) **Condition1**: One value is 2522 which is very high than others.
10) **Condition2**: One value is 2900 which is very high than others.
11) **Bldg Type**: One value is 2425 which is very high than others.
12) **Roof Style**: One value is 2321 which is very high than others.
13) **Roof Matl**: One value is 2887 which is very high than others.
14) **Mas Vnr Area**: Min is 0, Max is 1600 with median 0.
15) **Mas Vnr Type**: None has max values (1752).
16) **Exter Cond**: One value is 2549 which is very high than others.
17) **Bsmt Cond**: One value is 2616 which is very high than others.
18) **BsmtFin Type 2**: One value is 2499 which is very high than others.
19) **BsmtFin.SF.2**: Min is 0, max is 1526 with median 0.
20) **Heating**: One value is 2885 which is very high than others.
21) **Central Air**: One value is 2734 which is very high than others.
22) **Electrical**: One value is 2682 which is very high than others.
23) **X2nd.Flr.SF**: Min is 0, max is 2065 with median 0.
24) **Low Qual Fin SF**: Min is 0, max is 1064 and median is 0.
25) **Functional**: One value is 2728 which is very high than others.
26) **Fireplaces**: Most of the values are 1's.
27) **Fireplace Qu**: There are 1422 NA's.
28) **Garage Qual**: This attributes was not considered since 2615 out of 2930 entries have the 'Garage.Qual' as 'TA'.
29) **Garage Cond**: This attributes was not considered since 2665 out of 2930 entries have the 'Garage.Cond' as 'TA'.
30) **Paved Drive**: This attributes was not considered since 2652 out of 2930 entries have the 'Paved.Drive' as 'Y'.
31) **Wood Deck SF**: Since the median for this list of values is 0.0 when the Max value is 1424.00 and the mean is 93.75
32) **Open Porch SF**: Since the median for this list of values is 27.0 when the Max value is 742.00 and the mean is 47.53
33) **Enclosed Porch**: Since the median for this list of values is 0.0 when the Max value is 23.00 and the mean is 1012.00
34) **X3Ssn Porch**: Since the median for this list of values is 0.0 when the Max value is 508.00 and the mean is 2.592

35) **Screen Porch**: Since the median for this list of values is 0.0 when the Max value is 576 and the mean is 16
36) **Pool Area**: Since the median for this list of values is 0.0 when the Max value is 800.00 and the mean is 2.243
37) **Pool QC**: The number of entries marked N/A is 2917
38) **Fence**: The number of entries marked N/A is 2358
39) **Misc Feature**: The number of entries marked N/A is 2824
40) **Misc Val**: Since the median for this list of values was 0.0 when the Max value is 17000.00 and the mean is 50.63
41) **Mo Sold**: The month would not affect the sale price.
42) **Sale Type**: This attributes was not considered since 2536 out of 2930 entries have the 'Sale.Type' as 'WD'.
43) **Sale Condition**: This attributes was not considered since 2413 out of 2930 entries have the 'Sale.Condition' as 'Normal'.

We have then omitted the NA values followed by removing houses with more than 4000 sq. ft. which were outliers. Finally, the number of instances used for building the models were 2231.

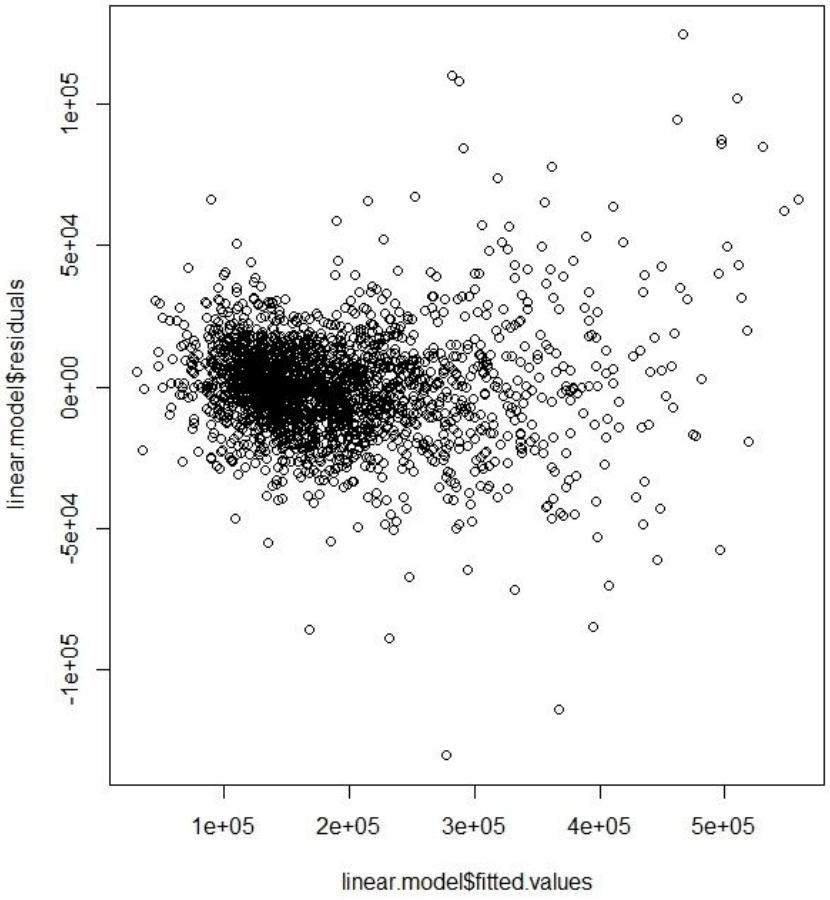
2) Models:

a) Linear Model:

i) A Simple Linear Regression model was built on the cleaned dataset with the value to be predicted as 'SalePrice' and this is done using the rest of the 38 attributes in the dataset. The plot of the "fitted_values" against "residuals" of the model was fairly linear, as shown in Fig. 1, stating that no further works needs to be done on it.

ii) After the model was built, the Multiplied R-squared value is 0.9494, Adjusted R-squared value is 0.9368, F-statistic: 75.29 on 445 and 1785 degrees of Freedom and a p-value which is less than $2.2\ e^{-16}$.

iii) The dataset was Cross-Validated using the K-fold method. The delta[1], i.e.,MSE value obtained is 521657318.



iv)

Fig.1: Plot of Linear Regression model

b) Forward and Backward Subset Selection:

i) Regression subsets were selected from the original dataset using the regsubsets() method of the leaps package.

ii) The **forward** subset was formed by selecting the forward option in the method attribute of regsubsets(). Also, the maximum size of the subset to be examined is 38, which is the number of attributes in the dataset. When a graph of 'number of variables' is plotted against the 'Adjusted R Square' values, the suitable number of variables for the subset looks like 8, guessed using Elbow Method. The graph obtained for the "Adjusted R squares" of the forward model is in Fig. 2.

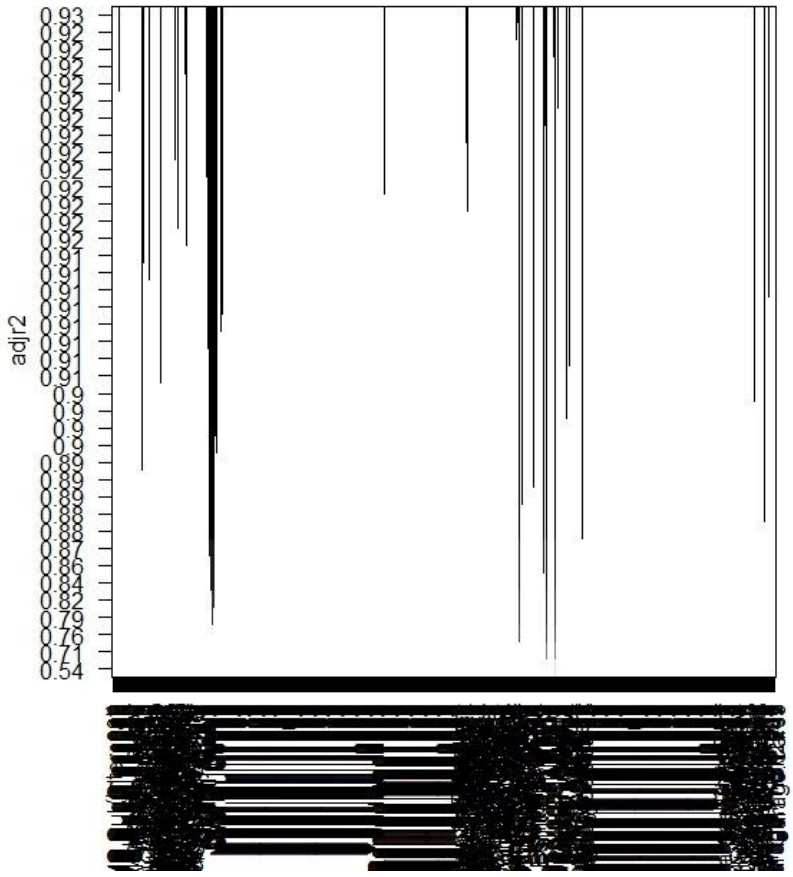


Fig. 2: "Adjusted R Squares" of Forward Subset

iii) The **backward** subset was formed by selecting the backward option in the method attribute of regsubsets(). Also, the maximum size of the subset to be examined is 58, which is the number of attributes in the dataset. When a graph of 'number of variables' is plotted against the 'Adjusted R Square' values, the suitable number of variables for the subset looks like 9, guessed using Elbow Method. The graph obtained for the "Adjusted R squares" of the backward model is in Fig. 3.

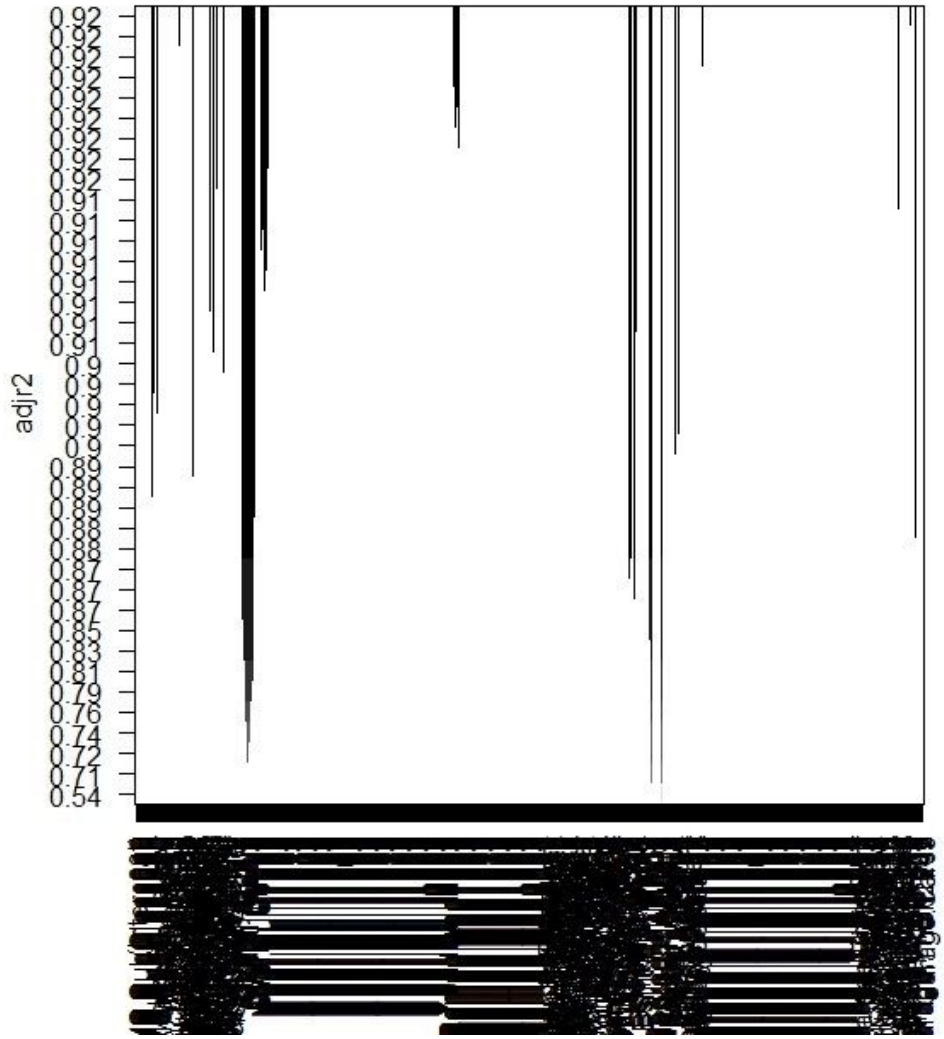


Fig. 3: "Adjusted R Squares" of Backward Subset

c) Building Ridge and Lasso Models:

To build the Ridge and Lasso models, we use the "glmnet" package. The glmnet(x, y) has two arguments. The first one loads an input matrix and the latter loads the response vector 'SalePrice' from the AmesHousing dataset. The function 'glmnet' gives a set of models for the users to choose from. Generally, for huge datasets, it is not easy for user to select a model. So, we use Cross Validation for this purpose.

In the command 'cv.out=cv.glmnet(x[train,],y[train],alpha=0)' , cv.glmnet returns an object, in this case it is cv.out which is a list of all the cross validation fits. The plot for cv.out is shown in the Fig. 4 below.

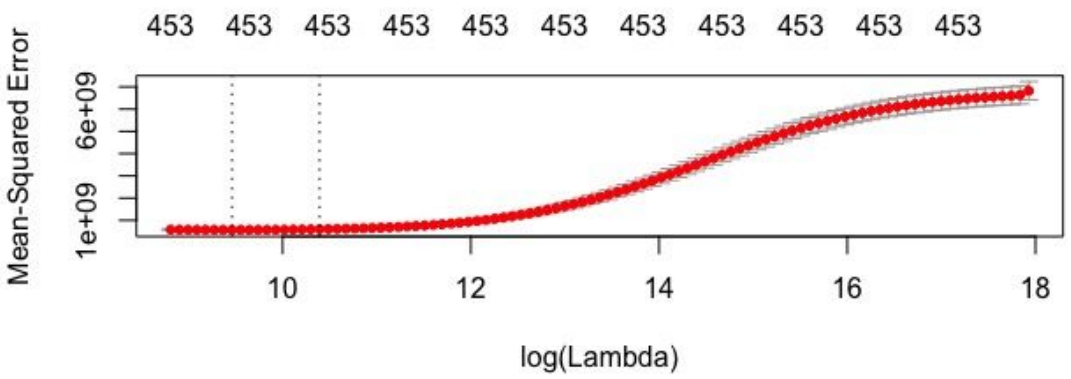


Fig. 4 Cross Validation Curve for Ridge model

In the above figure, red dotted line includes the cross-validation curve, and upper and lower standard deviation curves along the λ sequence. The two selected λ's are indicated by the vertical dotted lines. The test MSE for Ridge model obtained was '532852361'.

We then built the Lasso model using the same procedure except for a change in the value of alpha. We then do the Cross Validation similar to the Ridge model using the command 'cv.out =cv.glmnet (x[train ,],y[train],alpha =1)' and plot the cv.out to get the Fig. 5 below.

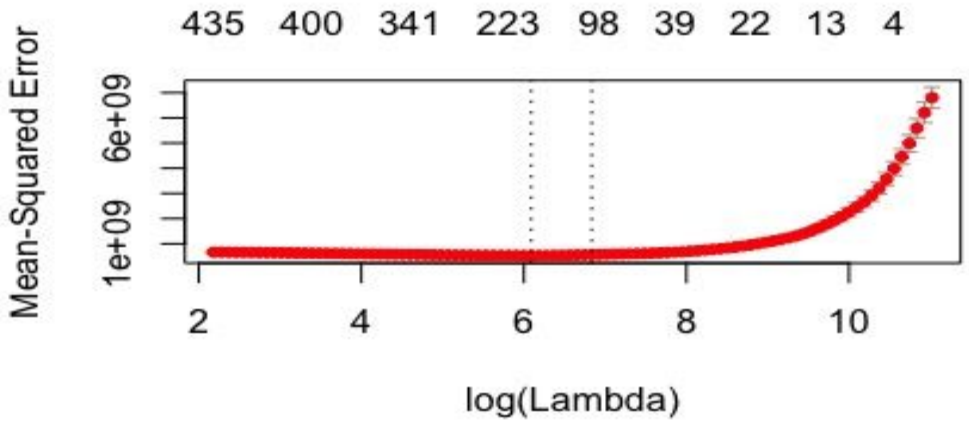


Fig. 5 Cross Validation Curve for Lasso model

The test MSE for Lasso model obtained was '505328176' which is similar to what we obtained in Ridge model.

3) Summary:

a) The total number of attributes being 82, we have removed 42 as they are not predictive. The attribute to be predicted is 'SalePrice'. The models were built using the remaining 39 predictors.

b) The test MSEs for Ridge and Lasso models obtained were very much similar.

c) The SQRT(MSE) / Avg SalePrice obtained for each method are as follows:

   i) Linear: 12.2493

   ii) Ridge: 12.4039

   iii) Lasso: 12.0794

Since the values are pretty similar to each other, it is very difficult to judge which the best model is but Lasso it is for its least value.